Lecture 23: Linear Regression Example & Multiple Linear Regression

STAT 324

Ralph Trane
University of Wisconsin–Madison

Spring 2020

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Linear Regression

Question: can we accurately estimate body fat percentage based on wrist circumference?

This would be excellent: body fat percentage is a good proxy for certain health statuses, but also difficult to measure. If we can estimate body fat percentage based on an easy to find measure such as wrist circumference, that would be great!

We will try to see if we can use a linear regression model to do so. I.e. we want to explain the relationship between body fat percentage and wrist circumference as

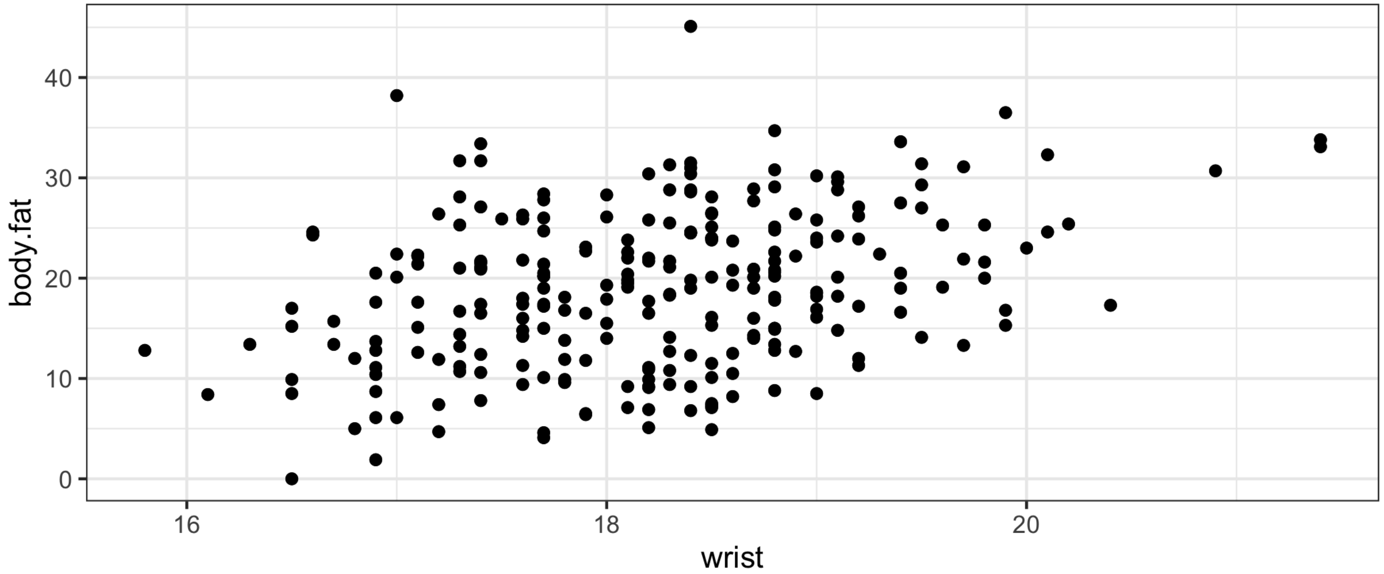
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim_{iid} N(0, \sigma^2).$$

Assumptions:

1. Relationship is linear
2. Observations independent
3. Variation around straight line is constant
4. Variation around straight line follows a normal distribution with mean 0

Assumption 1: linearity

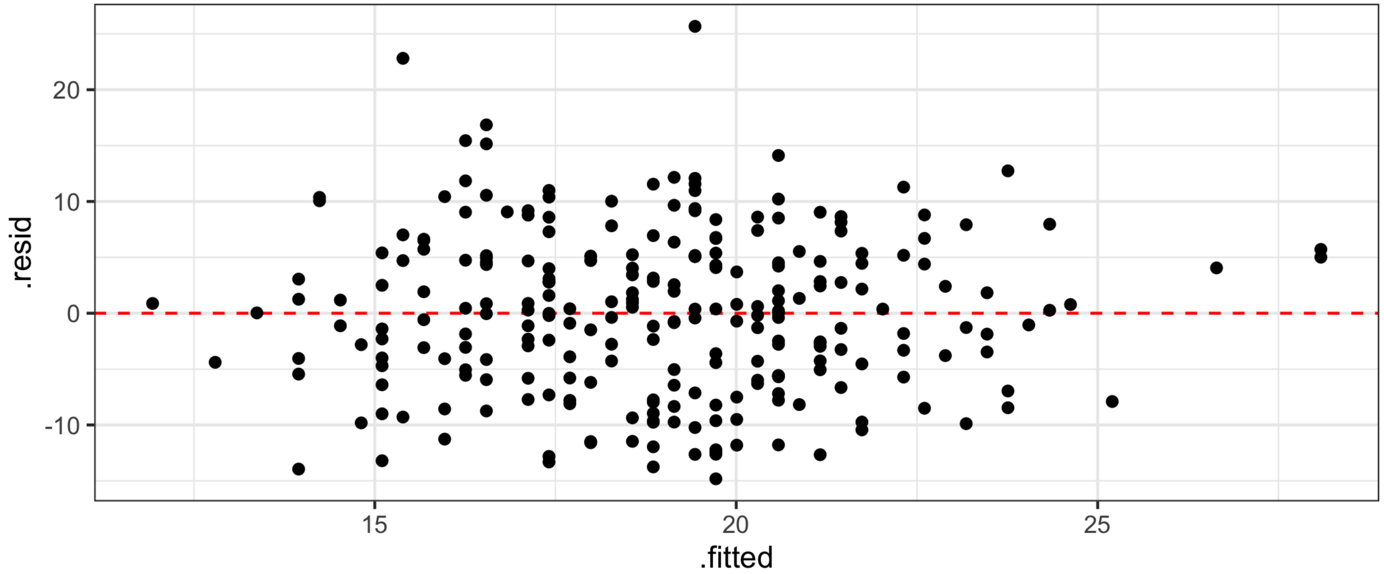
```
ggplot(body_fat,
       aes(x = wrist, y = body.fat)) +
  geom_point() +
  geom_point()
```



Linear Regression

Assumptions 2 and 3: independence and constant variance.

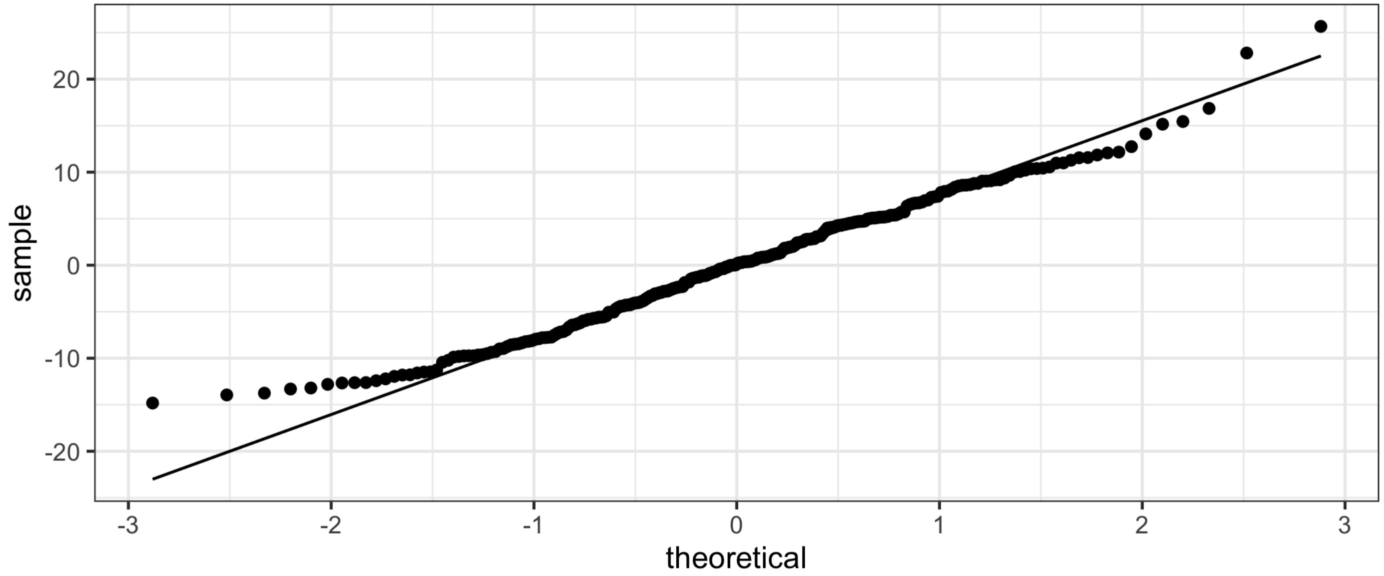
```r
lin_mod <- lm(data = body_fat, body.fat ~ wrist)

library(broom)
ggplot(augment(lin_mod),
       aes(x = .fitted, y = .resid)) +
  geom_hline(yintercept = 0, linetype = "dashed",
             color = "red") +
  geom_point()
```



Linear Regression

Assumption 4: normality

```
ggplot(augment(lin_mod),
       aes(sample = .resid)) +
  geom_qq() + geom_qq_line()
```



Linear Regression

```
summary(lin_mod)
```

```
Call:
lm(formula = body.fat ~ wrist, data = body_fat)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8183  -5.5840   0.1231   5.0703  25.6703 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -33.6660     8.9870  -3.746 0.000223 ***
wrist         2.8856     0.4923   5.861 1.45e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.

Residual standard error: 7.282 on 250 degrees of freedo
Multiple R-squared:  0.1208,    Adjusted R-squared:  0.
F-statistic: 34.35 on 1 and 250 DF,  p-value: 1.446e-08
```

- Positive coefficient: suggests body fat percentage increases as wrist circumference increases
- Testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$ gives a very small p-value which would lead us to reject at any reasonable level of α
- Assuming the linear model is correct, about 12% of the variation of body fat percentage measurements can be accounted for by the wrist circumference measurements.

Linear Regression

Maybe a separate study had previously indicated that an increase in wrist circumference of 1cm leads to an increase of 3 percentage point in the expected body fat percentage.

Let's check if this is compatible with our data. I.e. we want to test $H_0 : \beta_1 = 3$ vs $H_A : \beta_1 \neq 3$. We will use $\alpha = 0.1$. To do so, we first find our observed test statistic:

$$T_{\text{obs}} = \frac{\hat{\beta}_1 - 3}{\widehat{\text{SD}}(\hat{\beta}_1)} = \frac{2.886 - 3}{0.492}$$

```
(tobs <- (2.886 - 3)/0.492)
```

```
[1] -0.2317073
```

IF the null hypothesis is true, then $T \sim t_{n-2}$. (Note: $n-2$ is also dfE, or the residual degrees of freedom.) So, p-value:

```
2*cdf(StudentsT(df = nrow(body_fat) - 2), tobs)
```

```
[1] 0.8169549
```

We would NOT reject the null hypothesis since the p-value is greater than $\alpha = 0.1$.

Introduction to Multiple Linear Regression

But why only use wrist circumference? Why not try to include more data/variables?

case	body.fat	body.fat.siri	density	age	weight	height	BMI	ffweight
1	12.6	12.3	1.0708	23	154.25	67.75	23.7	134.9
2	6.9	6.1	1.0853	22	173.25	72.25	23.4	161.3
3	24.6	25.3	1.0414	22	154	66.25	24.7	116
4	10.9	10.4	1.0751	26	184.75	72.25	24.9	164.7
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

The linear regression framework is easily extended to allow us to include more variables.

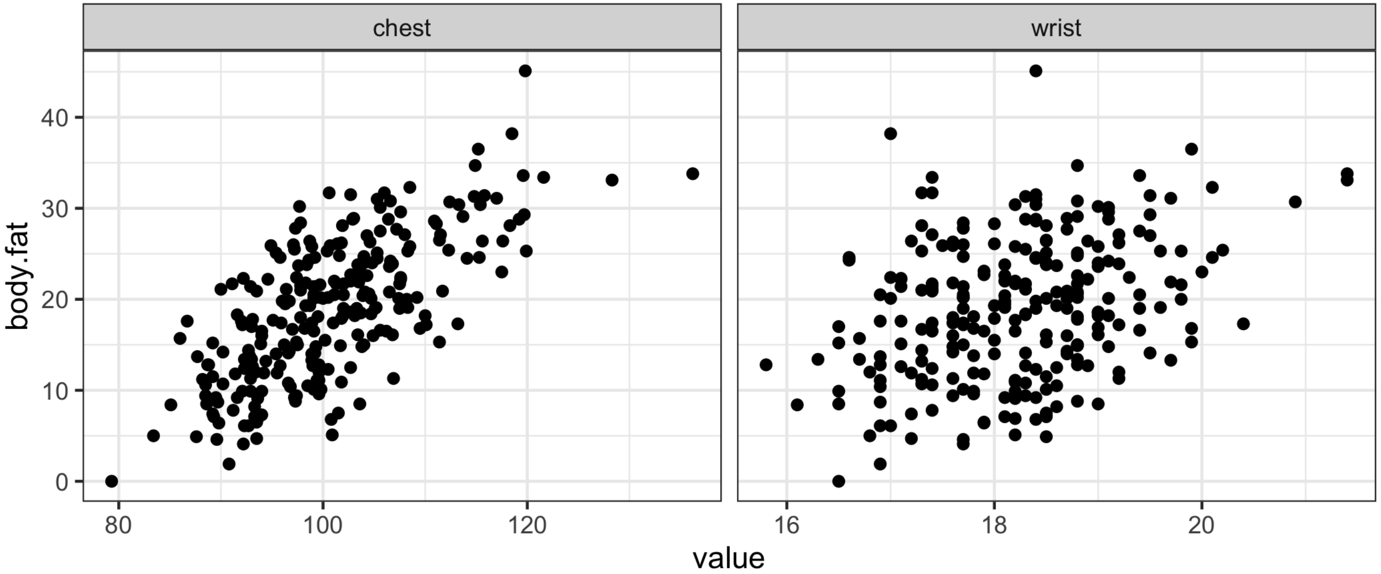
$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i, \quad \epsilon_i \sim_{iid} N(0, \sigma^2).$$

Assumptions:

1. Relationship really follows the equation provided above
2. Observations independent
3. Variation around straight line is constant
4. Variation around straight line follows a normal distribution with mean 0

Introduction to Multiple Linear Regression

Assumption 1: gets harder and harder, the more variables you include.



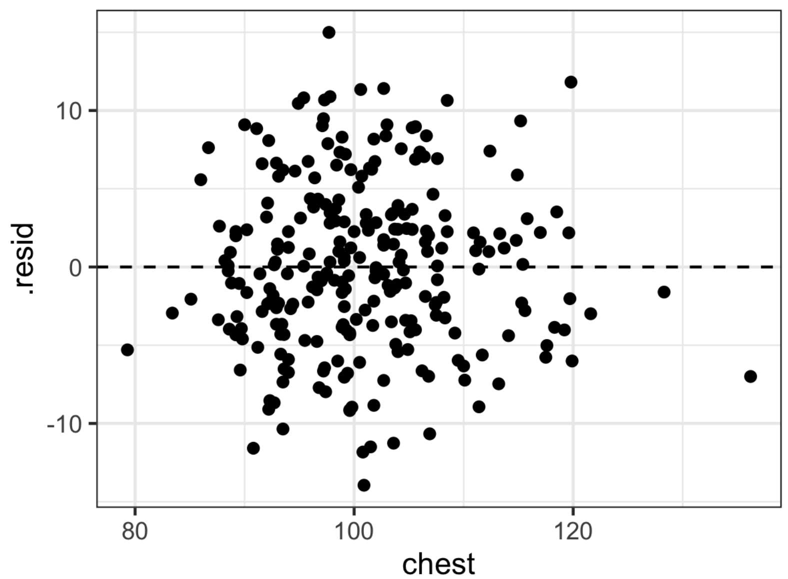
Introduction to Multiple Linear Regression

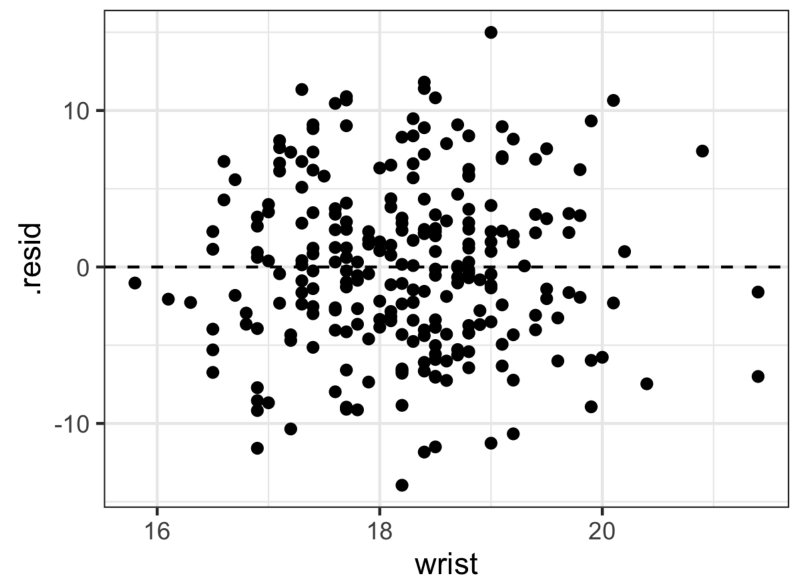
Assumption 2 and 3:

```
lin_mod2 <- lm(data = body_fat,
               body.fat ~ wrist + chest)
```

```
ggplot(augment(lin_mod2),
       aes(x = chest, y = .resid)) +
  geom_point() + geom_hline(yintercept =
```

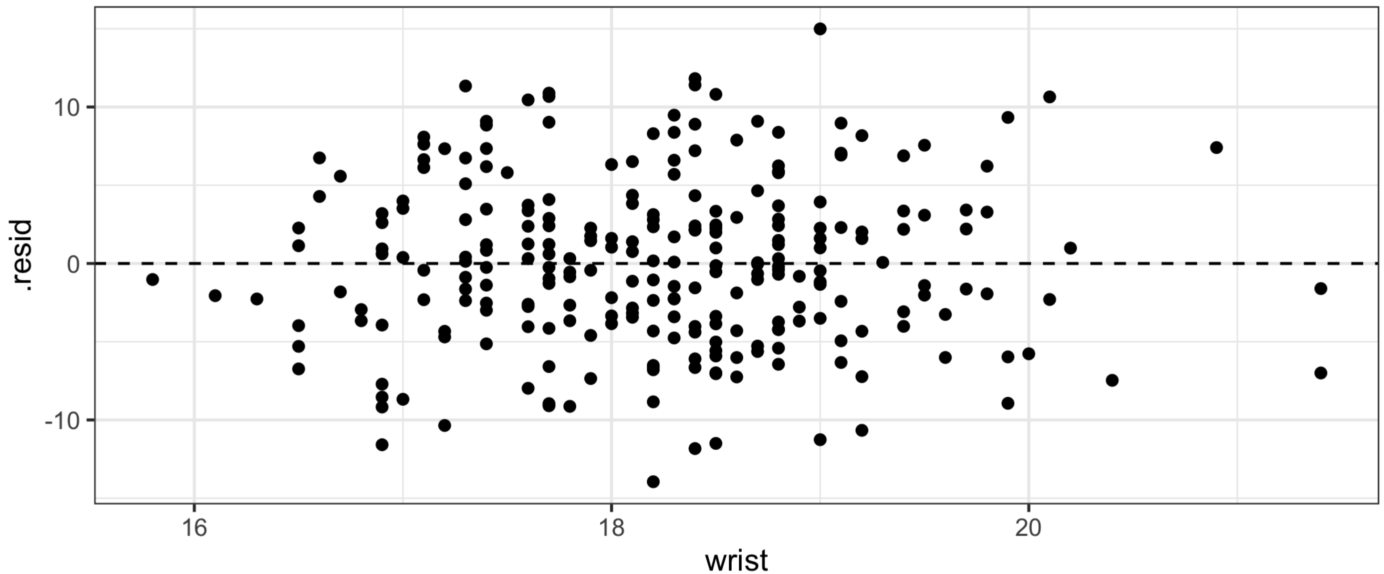
```
ggplot(augment(lin_mod2),
       aes(x = wrist, y = .resid)) +
  geom_point() + geom_hline(yintercept =
```





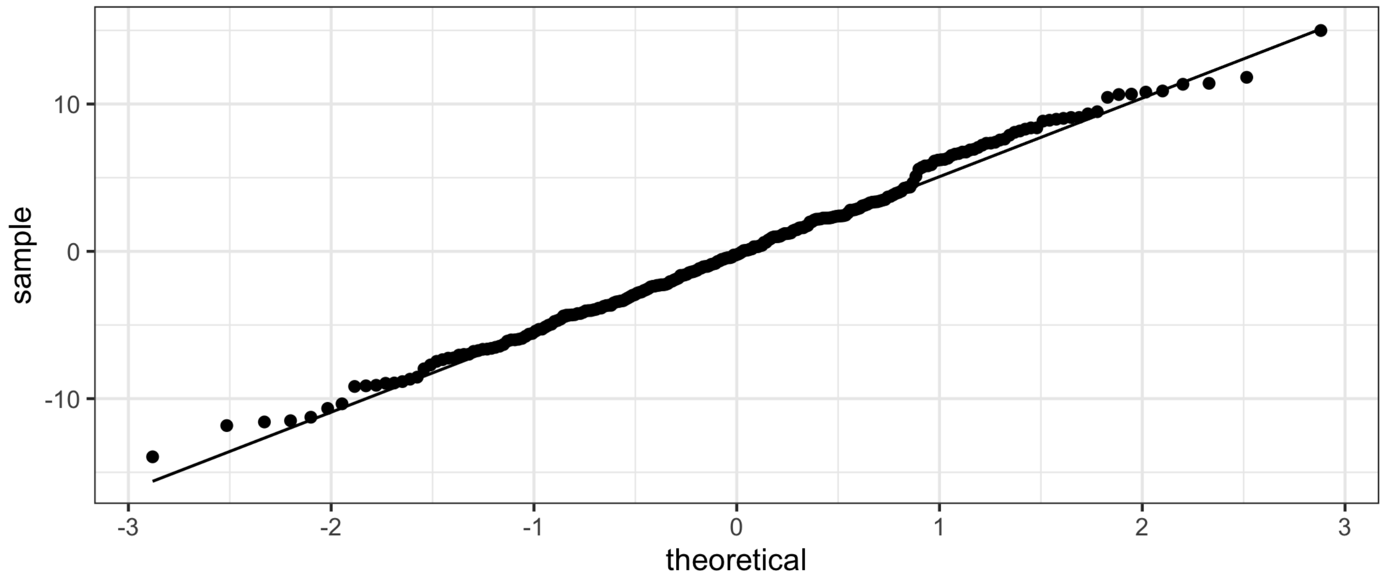
Assumption 2 and 3:

```r
ggplot(augment(lin_mod2),
       aes(x = wrist, y = .resid)) +
  geom_point() + geom_hline(yintercept = 0, linetype = "dashed")
```



Assumption 4: normality

```
ggplot(augment(lin_mod2),
       aes(sample = .resid)) +
  geom_qq() + geom_qq_line()
```



Introduction to Multiple Linear Regression

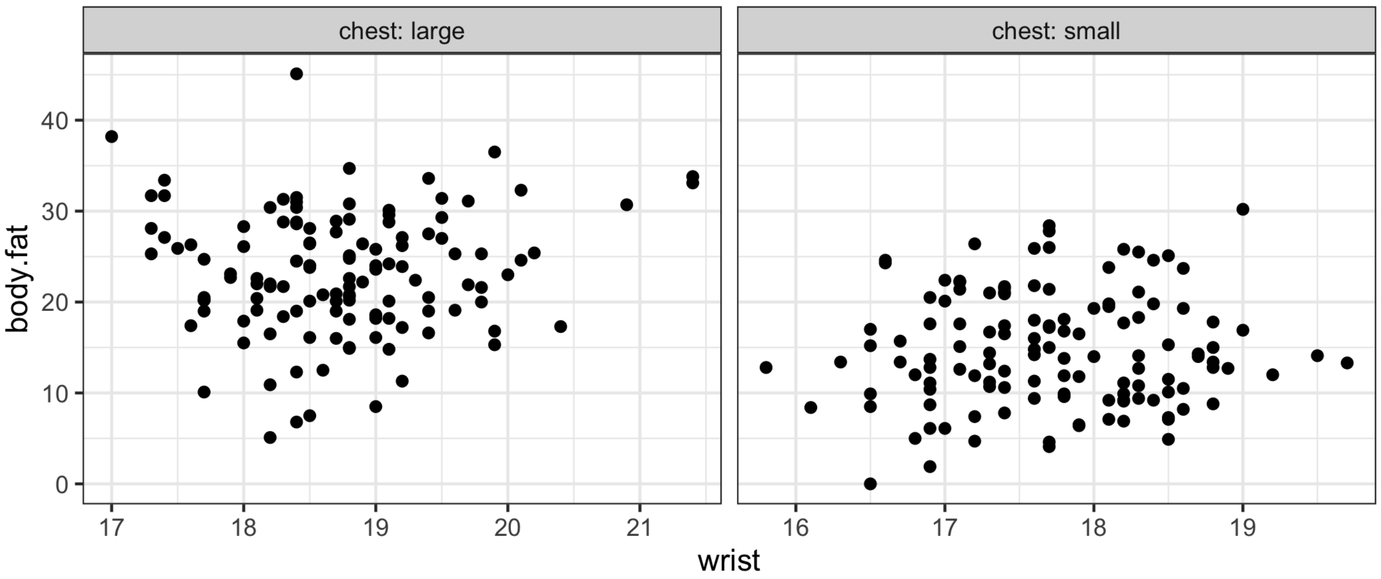
```
summary(lin_mod2)
```

```
Call:
lm(formula = body.fat ~ wrist + chest, data = body_fat)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.9480  -3.8502  -0.2248   3.3415  14.9917 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -27.60907    6.68039  -4.133  4.9e-05 ***
wrist        -1.71355    0.48626  -3.524 0.000506 ***
chest         0.77149    0.05385  14.327  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.

Residual standard error: 5.402 on 249 degrees of freedo
Multiple R-squared:  0.5181,    Adjusted R-squared:  0.
F-statistic: 133.8 on 2 and 249 DF,  p-value: < 2.2e-16
```

- chest: positive coefficient indicates body fat percentage increases with increased values of chest circumference *holding everything wirst constant*
- wrist: negative (!!) coefficient indicates body fat percentage decreases with increased values of wrist circumference *holding everything chest constant*
- both seems significant.

Introduction to Multiple Linear Regression

```r
body_fat %>%
  select(body.fat, wrist, chest) %>%
  mutate(chest = if_else(chest > median(chest), "large", "small")) %>%
  ggplot(aes(x = wrist, y = body.fat)) +
    geom_point() + facet_grid(~chest, scales = "free_x", labeller = label_both)
```



Introduction to Multiple Linear Regression

So, multiple linear regression allows us to utilize data from more than one variable. This is not without downsides:

- you have to worry about overfitting
- collinearity
- ...

Another reason why multiple linear regression is so important is the idea of adjusting for other covariates. An example:

In 1991, Radelet and Pierce wrote a paper titled "Choosing Those Who Will Die: Race and the Death Penalty in Florida".

Hypothesis: race influences how likely you are to receive the death penalty.

They obtained data on homicides and death sentences from Florida in the period 1976 through 1987.

defendant	death_penalty	no_death_penalty
white	53	430
black	15	176

Data from An Introduction to Categorical Data Analysis by Agresti (2007), originally from this paper.

Results you are more likely to receive the death penalty if you are white.

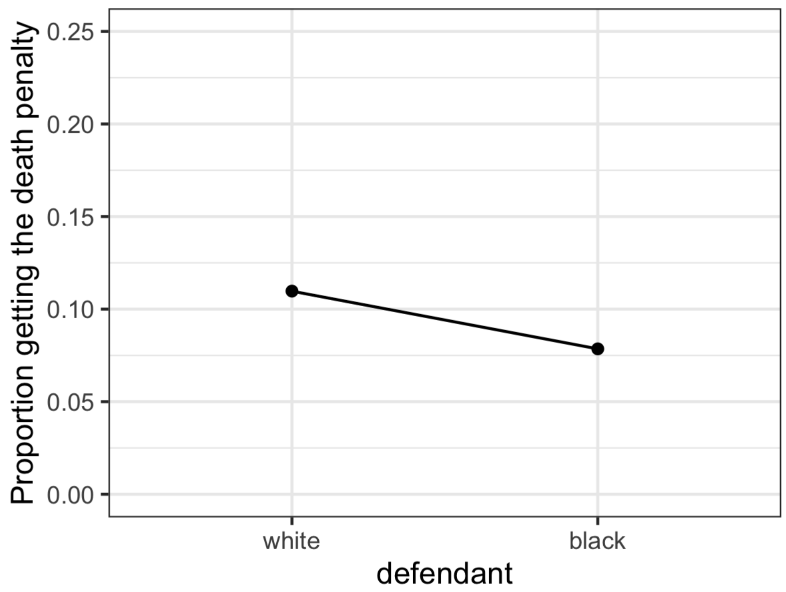
In fact, 10.97% of white defendants got the death penalty, while only 7.85% of black defendants received the same verdict.

This is of course not enough to convince you, so let's consider a test for difference in proportions and corresponding CI:

```r
prop.test(x = c(53, 15), n = c(53+430, 15+176), correct = FALSE)
```

```
    2-sample test for equality of proportions without continuity
    correction

data:  c(53, 15) out of c(53 + 430, 15 + 176)
X-squared = 1.4685, df = 1, p-value = 0.2256
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01605167  0.07844531
sample estimates:
    prop 1     prop 2
0.10973085 0.07853403
```

In a picture:

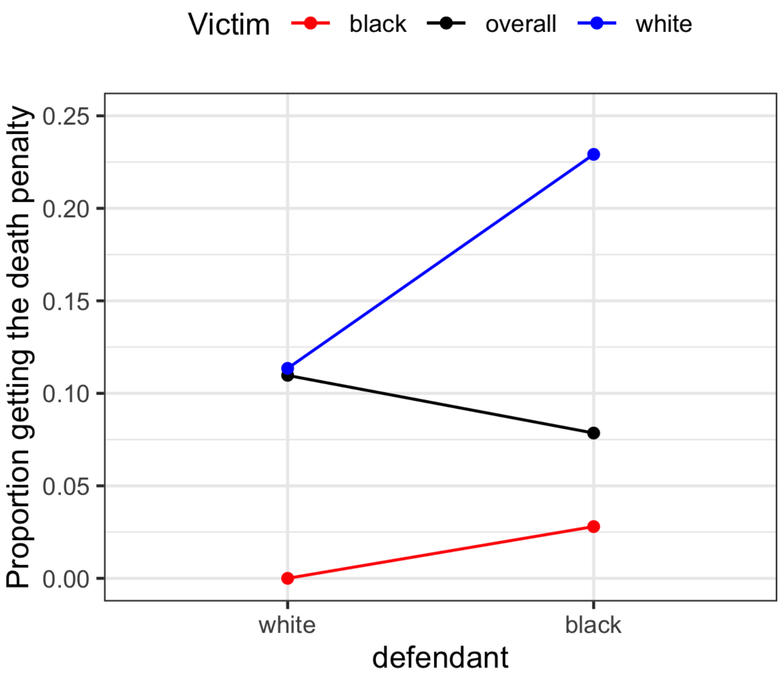


However, this is not the full story. There is one really strong confounding variable, which changes everything...

Race of the victim.

Introduction to Multiple Linear Regression

Once we adjust for race of the victim, things change dramatically:



Full data (counts)

defendant	victim	death_penalty	no_death_penalty	Total
white	white	53 (11.3%)	414 (88.7%)	467 (100.0%)
black	white	11 (22.9%)	37 (77.1%)	48 (100.0%)
white	black	0 (0.0%)	16 (100.0%)	16 (100.0%)
black	black	4 (2.8%)	139 (97.2%)	143 (100.0%)
Total	-	68 (10.1%)	606 (89.9%)	674 (100.0%)

Introduction to Multiple Linear Regression

This is a very simple example, where a regression might not be necessary, but the idea remains: there are situations where you need to correct for other variables. Regression provides a way to do just that.

Result of logistic regression *without* correction:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.093	0.146	-14.380	0.000	-2.390	-1.818
defendantblack	-0.369	0.306	-1.206	0.228	-1.001	0.207

Outcome: chance of receiving death penalty.

Interpretation of coefficients: if negative, chance decreases. If positive, chance increases.

Though not "statistically significant", it suggests black defendants less likely to receive death penalty, or at least not more likely.

Result of logistic regression *with* correction:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.059	0.146	-14.121	0.000	-2.356	-1.784
defendantblack	0.868	0.367	2.364	0.018	0.114	1.563
victimblack	-2.404	0.601	-4.004	0.000	-3.718	-1.307

Suggests (and achieves "statistical significance") that defendants who are black are *more* likely to achieve the death penalty, and if the victim was black, the defendant is *less* likely to achieve the death penalty.

...