Social Science Inquiry II

Week 8: Inference for multivariate regression, part I

Molly Offer-Westort

Department of Political Science,
University of Chicago

Winter 2022

Loading packages for this class

```
> library(ggplot2)
> library(estimatr)
> library(gridExtra)
> set.seed(60637)
```

- Housekeeping.

Recall our multivariate regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K \epsilon$$

Recall our multivariate regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K \epsilon$$

- How do we interpret β_0?

Recall our multivariate regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K \epsilon$$

- How do we interpret β_0? β_1?

Recall our multivariate regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K \epsilon$$

- How do we interpret β_0? β_1?β_K?

- We observe some data, which we (maybe) assume is randomly sampled from a larger population.

- We observe some data, which we (maybe) assume is randomly sampled from a larger population.
- The model describes the true relationships among the variables.

- We observe some data, which we (maybe) assume is randomly sampled from a larger population.
- The model describes the true relationships among the variables.
- But the true population parameters are generally unknown.

- We estimate the parameter values for a given sample, as the values that minimize the sum of squared residuals.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki}$$

- We estimate the parameter values for a given sample, as the values that minimize the sum of squared residuals.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki}$$

- Recall that the residual is defined as:

$$\hat{\epsilon}_i = \hat{Y}_i - Y_i$$

- We estimate the parameter values for a given sample, as the values that minimize the sum of squared residuals.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki}$$

- Recall that the residual is defined as:

$$\hat{\epsilon}_i = \hat{Y}_i - Y_i$$

- We will think about our random sample being not just for one variable, but from the joint distribution of $(Y, X_1, X_2, \ldots, X_K)$.

- We estimate the parameter values for a given sample, as the values that minimize the sum of squared residuals.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki}$$

- Recall that the residual is defined as:

$$\hat{\epsilon}_i = \hat{Y}_i - Y_i$$

- We will think about our random sample being not just for one variable, but from the joint distribution of $(Y, X_1, X_2, \ldots, X_K)$.
- Then each $\hat{\beta}_k$ is also random, with its own sampling distribution.

- We can get a point estimate for each of the parameters, $\hat{\beta}_k$: the coefficients in our linear model.

- We can get a point estimate for each of the parameters, $\hat{\beta}_k$: the coefficients in our linear model.
- We also want to get an estimate of the standard errors of the estimates, $\sqrt{\hat{\mathrm{Var}}[\hat{\beta}_k]}$, to describe how much we think these coefficients vary across samples.

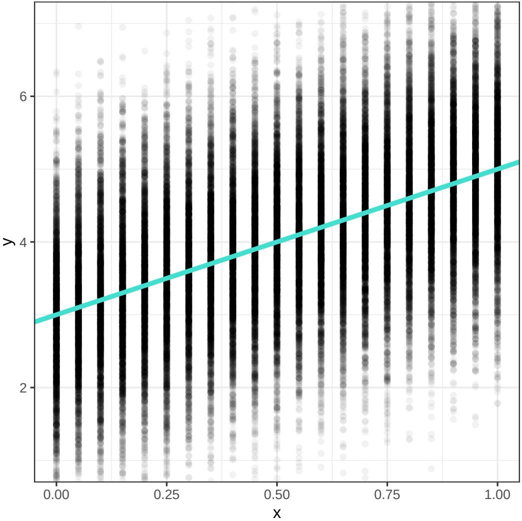
Suppose our true relationship is:

$$Y = 3 + 2X_1 + \epsilon$$

Suppose our true relationship is:

$$Y = 3 + 2X_1 + \epsilon$$

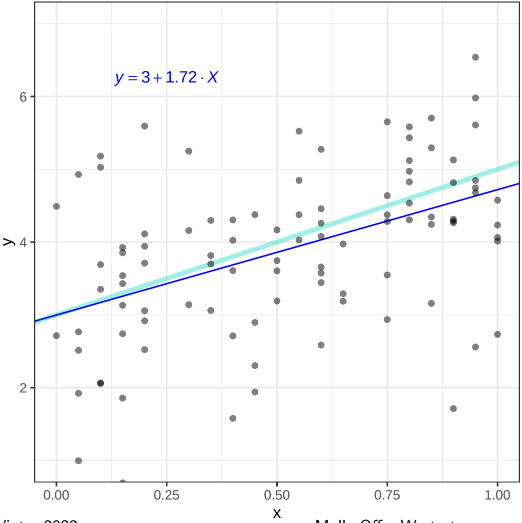
If we were to see the full data:



Suppose our true relationship is:

$$Y = 3 + 2X_1 + \epsilon$$

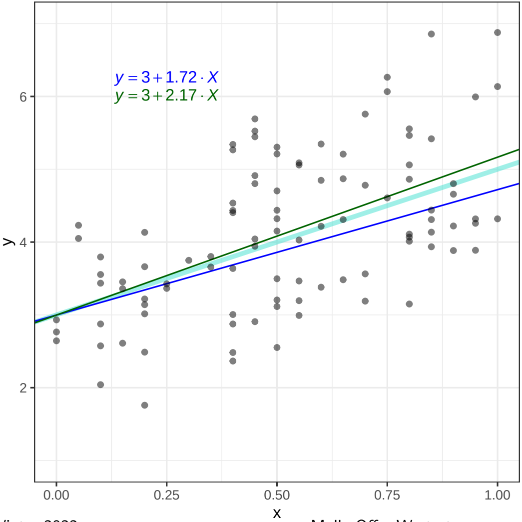
If we only saw 100 observations from the data:



Suppose our true relationship is:

$$Y = 3 + 2X_1 + \epsilon$$

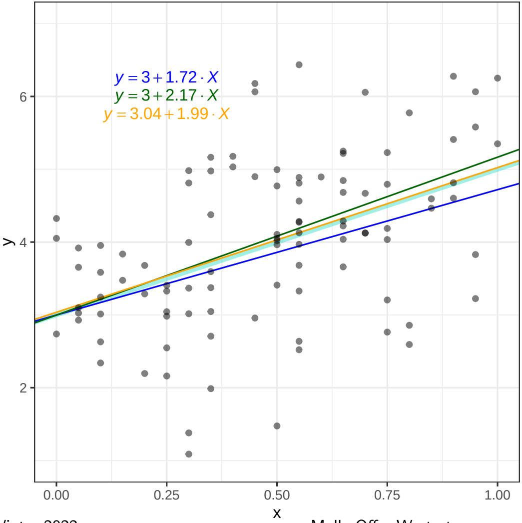
If we only saw 100 observations from the data:



Suppose our true relationship is:

$$Y = 3 + 2X_1 + \epsilon$$

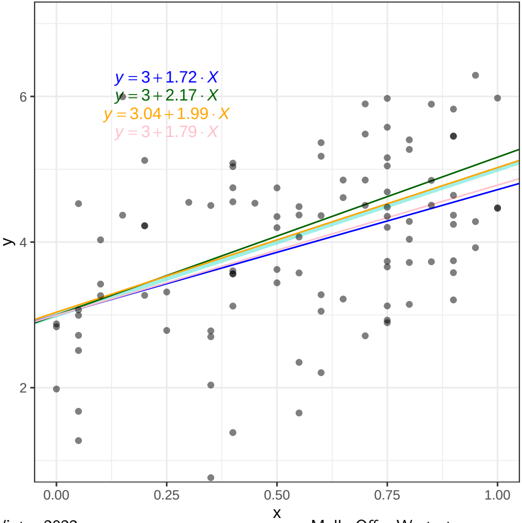
If we only saw 100 observations from the data:



Suppose our true relationship is:

$$Y = 3 + 2X_1 + \epsilon$$

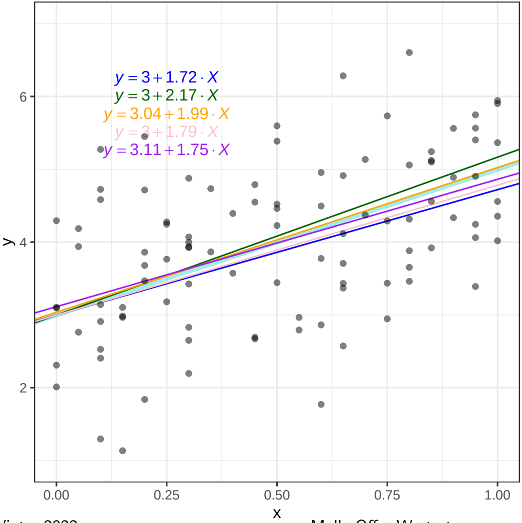
If we only saw 100 observations from the data:



Suppose our true relationship is:

$$Y = 3 + 2X_1 + \epsilon$$

If we only saw 100 observations from the data:



- Each time we sample from the population, we get a slightly different fit for our regression line.

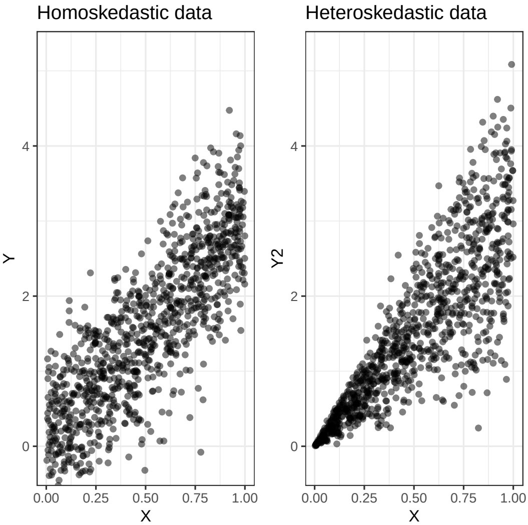
- Each time we sample from the population, we get a slightly different fit for our regression line.
- Our goal is to describe the *variability* in our parameter estimates.

- Each time we sample from the population, we get a slightly different fit for our regression line.
- Our goal is to describe the *variability* in our parameter estimates.
- Here, we have a regression line with just an intercept and a slope. But we could consider the same resampling and fitting procedure for any joint distribution of $(Y, X_1, X_2, \ldots, X_K)$,

- We will use **robust** standard errors.

- We will use **robust** standard errors.
- These standard errors don't require much beyond that our data is i.i.d.: random samples from the same joint distribution.

- We will use **robust** standard errors.
- These standard errors don't require much beyond that our data is i.i.d.: random samples from the same joint distribution.
- "Classical" regression modeling puts much stronger assumptions on the data, including that errors are "homoskedastic;" they don't vary with X


Homoskedastic data
Heteroskedastic data
Y
Y2
X
X
4
2
0
4
2
0
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00

Applied example

Recall:

Pager, D. (2003). The mark of a criminal record. *American Journal of Sociology*, 108(5), 937-975.

```
> dfp <- data.frame(
+   black = rep(c(0, 1), times = c(300, 400)),
+   record = c(rep(c(0, 1), each = 150),
+              rep(c(0, 1), each = 200)),
+   call_back = c(
+     # whites without criminal records
+     rep(c(0, 1), times = c(99, 51)), # 150
+     # whites with criminal records
+     rep(c(0, 1), times = c(125, 25)), # 150;
+     # - callbacks could be 25 or 26
+     # blacks without criminal records
+     rep(c(0, 1), times = c(172, 28)), # 200
+     # blacks with criminal records
+     rep(c(0, 1), times = c(190, 10)) # 200
+   )
+ )
>
```

- Let's try this with the dfp data, where the outcome Y is `call_back`, regressed on `black` and `record`, interacted.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Black}_i + \hat{\beta}_2 \text{Record}_i + \hat{\beta}_3 \text{Black}_i \times \text{Record}_i$$

```
> model2 <- lm_robust(call_back ~ black*record, data = dfp)
```

- How do we go about interpreting these coefficients? confidence intervals? p-values?

```
> summary(model2)

Call:
lm_robust(formula = call_back ~ black * record, data = dfp)

Standard error type:  HC2

Coefficients:
             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
(Intercept)    0.3400     0.0388    8.76 1.46e-17   0.2638   0.4162 696
black         -0.2000     0.0459   -4.35 1.55e-05  -0.2902  -0.1098 696
record        -0.1733     0.0494   -3.51 4.76e-04  -0.2703  -0.0764 696
black:record   0.0833     0.0573    1.45 1.46e-01  -0.0291   0.1958 696

Multiple R-squared:  0.0771 ,       Adjusted R-squared:  0.0732
F-statistic: 18.4 on 3 and 696 DF,  p-value: 1.76e-11
```

Confidence intervals

- A valid confidence interval CI_n for a target parameter θ with coverage $1 - \alpha$
$$\mathrm{P}[\theta \in CI_n] \geq 1 - \alpha$$

Confidence intervals

- A valid confidence interval CI_n for a target parameter θ with coverage $1 - \alpha$

$$\mathrm{P}[\theta \in CI_n] \geq 1 - \alpha$$

- If $\alpha = 0.05$, the probability that the estimand θ is in our confidence interval is greater than or equal to 0.95.

Confidence intervals

- A valid confidence interval CI_n for a target parameter θ with coverage $1 - \alpha$

$$\mathrm{P}[\theta \in CI_n] \geq 1 - \alpha$$

- If $\alpha = 0.05$, the probability that the estimand θ is in our confidence interval is greater than or equal to 0.95.
- CI_n is a random interval. It is a function of the data we observe.

Confidence intervals

- A valid confidence interval CI_n for a target parameter θ with coverage $1 - \alpha$
$$\mathrm{P}[\theta \in CI_n] \geq 1 - \alpha$$

- If $\alpha = 0.05$, the probability that the estimand θ is in our confidence interval is greater than or equal to 0.95.
- CI_n is a random interval. It is a function of the data we observe.
- θ is a fixed parameter. It does not move.

Confidence intervals

- A valid confidence interval CI_n for a target parameter θ with coverage $1 - \alpha$

$$\mathrm{P}[\theta \in CI_n] \geq 1 - \alpha$$

- If $\alpha = 0.05$, the probability that the estimand θ is in our confidence interval is greater than or equal to 0.95.
- CI_n is a random interval. It is a function of the data we observe.
- θ is a fixed parameter. It does not move.
(In the frequentist view of statistics.)

Confidence intervals

- A valid confidence interval CI_n for a target parameter θ with coverage $1 - \alpha$
$$\mathrm{P}[\theta \in CI_n] \geq 1 - \alpha$$
- If $\alpha = 0.05$, the probability that the estimand θ is in our confidence interval is greater than or equal to 0.95.
- CI_n is a random interval. It is a function of the data we observe.
- θ is a fixed parameter. It does not move.
(In the frequentist view of statistics.)
- If you use valid confidence repeatedly in your work, 95% of the time, your confidence intervals will include the true value of the relevant θ.

- The formula for the 95% confidence interval is:

$$CI_n = \left(\hat{\theta}_n - 1.96 \times \hat{\text{se}},\ \theta_n + 1.96 \times \hat{\text{se}}\right)$$

- The formula for the 95% confidence interval is:

$$CI_n = \left(\hat{\theta}_n - 1.96 \times \hat{\text{se}},\ \theta_n + 1.96 \times \hat{\text{se}}\right)$$

- The 1.96 value tells us how many standard errors away from the mean we need to include in our interval to get valid coverage.

- The formula for the 95% confidence interval is:

$$CI_n = \left(\hat{\theta}_n - 1.96 \times \hat{\mathrm{se}},\ \theta_n + 1.96 \times \hat{\mathrm{se}}\right)$$

- The 1.96 value tells us how many standard errors away from the mean we need to include in our interval to get valid coverage.
- This formula is based on a *normal approximation*, i.e., we assume the data is going to look like a normal distribution.

- How do we go about interpreting these coefficients? confidence intervals? p-values?

```
> confint(model2)

              2.5 % 97.5 %
(Intercept)   0.264  0.416
black        -0.290 -0.110
record       -0.270 -0.076
black:record -0.029  0.196
```

P-values

Suppose $\hat{\theta}$ is the general form for an estimate produced by our estimator, and $\hat{\theta}^*$ is the value we have actually observed.

P-values

- A two-tailed p-value under the null hypothesis is

$$p = \mathrm{P}_0[|\hat{\theta}| \geq |\hat{\theta}^*|]$$

i.e., the probability *under the null distribution* that we would see an estimate of $\hat{\theta}$ as or more extreme as what we saw from the data.

- Confidence intervals and p-values help us get back to testing hypotheses.

Two-sided hypotheses

$$H_0 : \theta = 0$$

$$H_A : \theta \neq 0$$

Two-sided hypotheses

$$H_0 : \theta = 0$$

$$H_A : \theta \neq 0$$

Note that we are *not* imposing the sharp null of no individual effect here, we're looking at averages.

P-values

- A two-tailed p-value under the null hypothesis is

$$p = \mathrm{P}_0[|\hat{\theta}| \geq |\hat{\theta}^*|]$$

i.e., the probability *under the null distribution* that we would see an estimate of $\hat{\theta}$ as or more extreme as what we saw from the data.

- Confidence intervals and hypothesis tests have a specific relationship.

- Confidence intervals and hypothesis tests have a specific relationship.
- Consider all of the hypotheses that take the form:

$$H_0 : \theta = \theta_0$$
$$H_A : \theta \neq \theta_0$$

- Confidence intervals and hypothesis tests have a specific relationship.
- Consider all of the hypotheses that take the form:

$$H_0 : \theta = \theta_0$$

$$H_A : \theta \neq \theta_0$$

- If the calculated two-tailed p-value is less than 0.05, reject the hypothesis.

- Confidence intervals and hypothesis tests have a specific relationship.
- Consider all of the hypotheses that take the form:

$$H_0 : \theta = \theta_0$$
$$H_A : \theta \neq \theta_0$$

- If the calculated two-tailed p-value is less than 0.05, reject the hypothesis.
- If the calculated two-tailed p-value is greater than 0.05, fail to reject the hypothesis.

- Confidence intervals and hypothesis tests have a specific relationship.
- Consider all of the hypotheses that take the form:

$$H_0 : \theta = \theta_0$$
$$H_A : \theta \neq \theta_0$$

- If the calculated two-tailed p-value is less than 0.05, reject the hypothesis.
- If the calculated two-tailed p-value is greater than 0.05, fail to reject the hypothesis.
- The θ_0 for which we would fail to reject the hypothesis lie within the 95% confidence interval.

One way that this is very useful:

One way that this is very useful:

- If 0 is outside the 95% confidence interval, we would reject the hypothesis that $\theta = 0$ at $p = 0.05$.

One way that this is very useful:

- If 0 is outside the 95% confidence interval, we would reject the hypothesis that $\theta = 0$ at $p = 0.05$.
- If 0 is outside the 99% confidence interval, we would reject the hypothesis that $\theta = 0$ at $p = 0.01$.

One way that this is very useful:

- If 0 is outside the 95% confidence interval, we would reject the hypothesis that $\theta = 0$ at $p = 0.05$.
- If 0 is outside the 99% confidence interval, we would reject the hypothesis that $\theta = 0$ at $p = 0.01$.
- If 0 is outside the 99.9% confidence interval, we would reject the hypothesis that $\theta = 0$ at $p = 0.001$.

- How do we go about interpreting these coefficients? confidence intervals? p-values?

```
> summary(model2)

Call:
lm_robust(formula = call_back ~ black * record, data = dfp)

Standard error type:  HC2

Coefficients:
             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
(Intercept)    0.3400     0.0388    8.76 1.46e-17   0.2638   0.4162 696
black         -0.2000     0.0459   -4.35 1.55e-05  -0.2902  -0.1098 696
record        -0.1733     0.0494   -3.51 4.76e-04  -0.2703  -0.0764 696
black:record   0.0833     0.0573    1.45 1.46e-01  -0.0291   0.1958 696

Multiple R-squared:  0.0771 ,       Adjusted R-squared:  0.0732
F-statistic: 18.4 on 3 and 696 DF,  p-value: 1.76e-11

> round(model2$p.value,5)

 (Intercept)        black       record black:record
     0.00000      0.00002      0.00048      0.14622
```

Bootstrap estimation

- Another approach to estimating the standard error of an estimate is to use bootstrapping.

Bootstrap estimation

- Another approach to estimating the standard error of an estimate is to use bootstrapping.
- If we fully knew the joint distribution of our population, we would know exactly how to determine the sampling variation of our estimate.

Bootstrap estimation

- Another approach to estimating the standard error of an estimate is to use bootstrapping.
- If we fully knew the joint distribution of our population, we would know exactly how to determine the sampling variation of our estimate.
- While we do not, we can *suppose* that the empirical joint distribution produced by the data that we observe is identical to the population joint distribution.

Bootstrap estimation

- Another approach to estimating the standard error of an estimate is to use bootstrapping.
- If we fully knew the joint distribution of our population, we would know exactly how to determine the sampling variation of our estimate.
- While we do not, we can *suppose* that the empirical joint distribution produced by the data that we observe is identical to the population joint distribution.
- We can then just re-sample with replacement from our observed data, and see how much our estimates vary across re-samples.

Bootstrap estimation

The bootstrapping procedure is:

- Repeat many times:
  1. Take a sample of size *n with replacement* from the observed data

Bootstrap estimation

The bootstrapping procedure is:

- Repeat many times:
  1. Take a sample of size *n with replacement* from the observed data
  2. Apply the estimating procedure on the bootstrap sample.

Bootstrap estimation

The bootstrapping procedure is:

- Repeat many times:
  1. Take a sample of size *n with replacement* from the observed data
  2. Apply the estimating procedure on the bootstrap sample.
- Calculate the standard deviation of a parameter estimate across these many bootstrap estimates.

We can try this with the Pager (2003) data.

```
> outmat <- replicate(1000, # do this 1000 times
+                     {
+                       # Take a sample of size n with replacemetn from the data
+                       idx <- sample(1:nrow(dfp), replace = TRUE)
+                       # fit the model on the sampled data
+                       lmx <- lm_robust(call_back ~ black*record,
+                                        data = dfp[idx,])
+                       coef(lmx)
+                     })
> outmat <- t(outmat)
> dim(outmat)
```

```
[1] 1000    4
```

```
> head(outmat, 4)
```

```
     (Intercept) black record black:record
[1,]        0.30 -0.17  -0.12        0.057
[2,]        0.45 -0.32  -0.22        0.136
[3,]        0.33 -0.18  -0.19        0.078
[4,]        0.34 -0.18  -0.16        0.051
```

We can try this with the Pager (2003) data.

```
> apply(outmat, 2, sd)

 (Intercept)        black       record black:record
      0.039        0.046        0.049        0.057
```

We can try this with the Pager (2003) data.

```
> apply(outmat, 2, sd)
```

```
 (Intercept)        black       record black:record
       0.039        0.046        0.049        0.057
```

Compare this to the robust standard errors from our model.

```
> model2$std.error
```

```
 (Intercept)        black       record black:record
       0.039        0.046        0.049        0.057
```

We can also get confidence intervals from the bootstrap estimates.

We can also get confidence intervals from the bootstrap estimates. For each parameter...

1. Sort bootstrap estimates from smallest to largest
2. Find the lower bound as the $\alpha/2$ percentile, and the upper bound$1 - \alpha/2$ percentile; i.e., so that $(1 - \alpha)\%$ of estimates are within this range

```
> t(apply(outmat, 2, quantile, probs = c(0.025, 0.075)))

              2.5%    7.5%
(Intercept)   0.27  0.2867
black        -0.29 -0.2681
record       -0.27 -0.2455
black:record -0.03  0.0027
```

We can also get confidence intervals from the bootstrap estimates. For each parameter. . .

1. Sort bootstrap estimates from smallest to largest
2. Find the lower bound as the $\alpha/2$ percentile, and the upper bound$1 - \alpha/2$ percentile; i.e., so that $(1 - \alpha)\%$ of estimates are within this range

```
> t(apply(outmat, 2, quantile, probs = c(0.025, 0.075)))
```

```
              2.5%    7.5%
(Intercept)   0.27  0.2867
black        -0.29 -0.2681
record       -0.27 -0.2455
black:record -0.03  0.0027
```

Compare this to the confidence intervals from our model.

```
> confint(model2)
```

```
              2.5 % 97.5 %
(Intercept)   0.264  0.416
black        -0.290 -0.110
record       -0.270 -0.076
black:record -0.029  0.196
```

- `estimatr::lm_robust()` outputs robust standard errors by default; this is why it's really nice to use.

- `estimatr::lm_robust()` outputs robust standard errors by default; this is why it's really nice to use.
- There are options for different types of robust standard errors which have different small sample properties, but they're asymptotically equivalent.

References I

Pager, D. (2003). The mark of a criminal record. American journal of sociology, 108(5):937–975.