Uncertainty, 2

Frank Edwards

We can describe the relationship between a predictor variable *X* and an outcome variable *Y* with the line:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Where β_0 is the y-intercept of the line, β_1 is the slope of the line, and ε is the error between the fitted line and the coordinates (X, Y)

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of *y* when *x* is equal to zero

β_1: The average increase in *y* when *x* increases by one unit

ε: The distance between the line $y = \beta_0 + \beta_1 X$ and the actual observed values of *y*. Allows us to estimate the line, even when x and y do not fall exactly on a line.

The line $y = \beta_0 + \beta_1 X$ provides a prediction for the values of *y* based on the values of *x*.

Remember, that we put a *hat* on variables to indicate that they are estimated from the data, or predicted.

In other words, we try to learn about the *regression coefficients* β_1 and β_0 by estimating $\hat{\beta}_1$ and $\hat{\beta}_0$.

A regression line predicts values Y, $\hat{Y}$ with the equation:

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

and the residual, or prediction error is the difference between the observed and predicted values of Y

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

Understanding the regression line
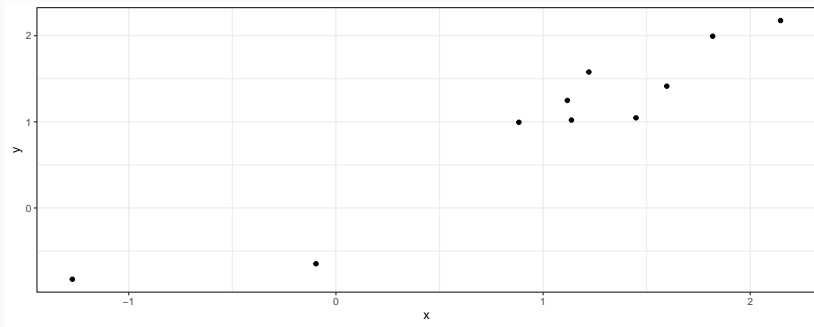
```
## # A tibble: 10 x 2
##          x       y
##      <dbl>   <dbl>
##  1  1.14    1.02
##  2  1.60    1.41
##  3  0.883   0.995
##  4  2.15    2.18
##  5  1.22    1.58
##  6  1.12    1.25
##  7  1.82    1.99
##  8 -0.0961 -0.648
##  9  1.45    1.05
## 10 -1.27   -0.828
```

$\hat{\beta}_0 = 0.05$, $\hat{\beta}_1 = 0.95$

- Estimate $\hat{Y}$. Recall that $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
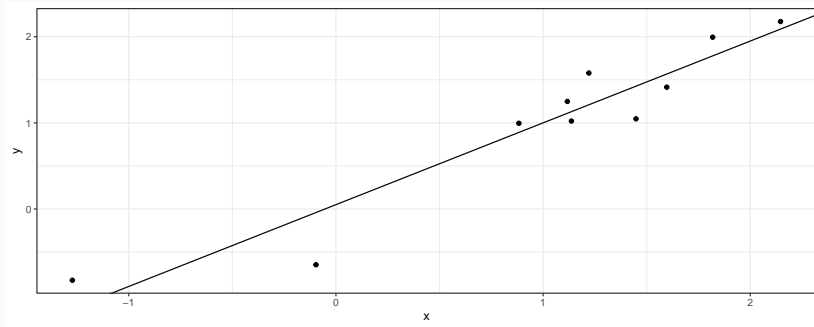- Estimate ε. Recall that $\varepsilon = Y - \hat{Y}$

$\hat{\beta}_0 = 0.05, \hat{\beta}_1 = 0.95$

$\hat{\beta}_0 = 0.05, \hat{\beta}_1 = 0.95$

$\hat{\beta}_0 = 0.05$, $\hat{\beta}_1 = 0.95$

$\hat{\beta}_0 = 0.05, \hat{\beta}_1 = 0.95$
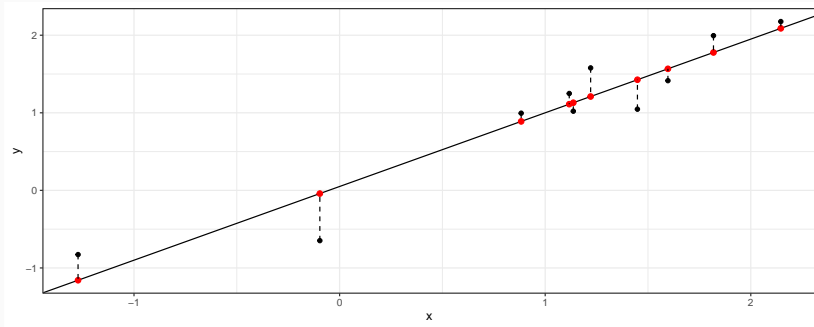
Goal of linear regression

1. Estimate causal relationships between some predictor variable *x* and outcome variable *y*
2. Predict changes in some outcome variable *y* for changes in a system of predictors *X*

For estimates of β to be unbiased on consistent, the following assumptions must be met:

1. The linear model approximates the data generating process
2. Exogeneity of errors: $E(\varepsilon) = 0$, errors uncorrelated with predictors (ALMOST ALWAYS THE CASE IN OBSERVATIONAL DATA)
3. Linear independence of predictors
4. Constant error variance (Homoskedasticity): $V(\varepsilon|X) = V(\varepsilon)$

For predictions $\hat{y} = \beta X$ to be unbiased and consistent, the following assumptions must be met

1. The linear model approximates the data generating process
2. Constant error variance (Homoskedasticity): $V(\varepsilon|X) = V(\varepsilon)$

1. When you cannot meet the exogeneity assumption (unmeasured confounding, no randomization) or the linear independence assumption, *you cannot interpret β as a causal estimate.*
2. When you cannot meet the assumption of a linear model or constant error variance, *you cannot make valid predictions*

As linear with Normal errors:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

As Normal, with linear vector of means:

$$y \sim N(\beta X, \sigma^2)$$

Example: OLS for estimating causal effects

```r
### read and format Pager data
cr <- read_csv("./data/criminalrecord.csv")
cr <- cr %>% mutate(race = case_when(black == 0 ~ "White", black == 1 ~ "Black"),
    crimrec = as.logical(crimrec)) %>% select(callback, race, crimrec)

head(cr)
```

```
## # A tibble: 6 x 3
##   callback race  crimrec
##      <dbl> <chr> <lgl>
## 1        1 White TRUE
## 2        0 White FALSE
## 3        1 White FALSE
## 4        1 White FALSE
## 5        0 White TRUE
## 6        0 White TRUE
```

```r
treatment <- cr %>% filter(crimrec == T) %>% group_by(race) %>% summarise(xbar = sum(callback)/n(),
    se = sd(callback)/sqrt(n()))

control <- cr %>% filter(crimrec == F) %>% group_by(race) %>% summarise(xbar = sum(callback)/n(),
    se = sd(callback)/sqrt(n()))

SATE <- treatment %>% mutate(SATE = xbar - control$xbar, SATE_se = sqrt(se^2 + control$se^2))
```

For estimates of β to be unbiased on consistent, the following assumptions must be met:

1. The linear model approximates the data generating process:
   Questionable under binary outcome, but ok
2. Exogeneity of errors: $E(\varepsilon) = 0$, errors uncorrelated with predictors
   Randomization handles this
3. Linear independence of predictors *Randomization also handles this*
4. Constant error variance (Homoskedasticity): $V(\varepsilon|X) = V(\varepsilon)$ *We can generally check this after estimation, but will be violated with binary outcome*

If we assume that the errors are Normally distributed with constant variance: $\varepsilon \sim N(0, \sigma^2)$, then we can treat the standard error of β as the standard deviation of its sampling distribution.

In other words: $\beta \sim N(\hat{\beta}, SE_\beta^2)$

If we assume that the errors are Normally distributed with constant variance: $\varepsilon \sim N(0, \sigma^2)$, then we can treat the standard error of β as the standard deviation of its sampling distribution.

In other words: $\beta \sim N(\hat{\beta}, SE_\beta^2)$

The standard error of β is calculated as:

$$SE_\beta = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Using the central limit theorem to calculate confidence intervals, compute p-values

If the sampling distribution for β is defined as:

$$\beta \sim N(\hat{\beta}, SE_\beta^2)$$

Then we can construct a CI for β with a critical value of α as:

$$\hat{\beta} \pm z_{\alpha/2} \times SE_\beta$$

And conduct a *z* test for β by comparing against the null hypothesis:

$$H_0 : \beta \sim N(0, SE_\beta^2)$$

```
cr_ols <- lm(callback ~ race * crimrec, data = cr)

library(broom)
library(pander)
pander(tidy(cr_ols))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.1407	0.02529	5.565	3.76e-08
raceWhite	0.1993	0.03857	5.167	3.112e-07
crimrecTRUE	-0.08994	0.03585	-2.509	0.01234
raceWhite:crimrecTRUE	-0.08339	0.0546	-1.527	0.1272

```r
pred_dat <- data.frame(race = c("Black", "White"), crimrec = c(T, F)) %>% complete(race,
    crimrec)
# complete() makes all possible combinations of listed variables

cr_ols_yhat <- as.data.frame(predict(cr_ols, newdata = pred_dat, interval = "confidence"))

pred_dat <- pred_dat %>% mutate(yhat = cr_ols_yhat$fit, yhat_lwr = cr_ols_yhat$lwr,
    yhat_upr = cr_ols_yhat$upr)
```

Check results

pander(pred_dat)

race	crimrec	yhat	yhat_lwr	yhat_upr
Black	FALSE	0.1407	0.09106	0.1903
Black	TRUE	0.05076	0.0008641	0.1007
White	FALSE	0.34	0.2828	0.3972
White	TRUE	0.1667	0.1095	0.2238

pander(treatment)

race	xbar	se
Black	0.05076	0.01568
White	0.1667	0.03053

pander(control)

race	xbar	se
Black	0.1407	0.02471
White	0.34	0.03881

Visualize results - with 95 percent confidence interval

```
ggplot(pred_dat, aes(x = race, y = yhat, ymin = yhat_lwr, ymax = yhat_upr, fill = crimrec)) +
    geom_crossbar(alpha = 0.7)
```
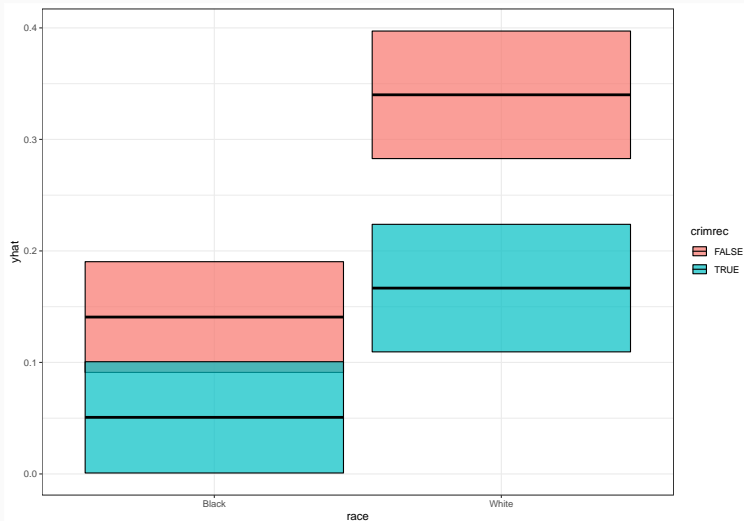
1. Randomized controlled trial
2. "Natural experiments": Difference in differences, other matched-group methods (propensity scores, etc)
3. Instrumental variables
4. Regression discontinuity

OLS for prediction and description

Police and municipal budgets

```
budgets <- read_csv("./data/police_spending.csv")
glimpse(budgets)
```

```
## Observations: 501
## Variables: 7
## $ fips                <chr> "01001", "01021", "01033", "01061", "01081", "0...
## $ exp_police_pc       <dbl> 169.24, 170.18, 192.94, 164.06, 186.56, 163.66,...
## $ officers_pc         <dbl> 83.36, 32.19, 31.90, 29.89, 51.58, 34.35, 19.80...
## $ rev_prop_tax_pc     <dbl> 158.17, 242.31, 256.04, 260.15, 330.72, 169.51,...
## $ violent.crime.high  <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, FALS...
## $ segregation.bw.high <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, ...
## $ segregation.lw.high <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, TRUE, TRUE, FA...
```

Spending on police per capita $= f($Property taxes $+$ Crime $+$ Segregation$)$

Develop a model to predict police department budgets

Spending on police per capita $= f(\text{Property taxes} + \text{Crime} + \text{Segregation})$

Using an OLS model:

$$y \sim N(\beta X, \sigma^2)$$

Where:

$$\beta X = \hat{y}_i = \beta_0 + \beta_1 \text{taxes}_i + \beta_2 \text{crime}_i + \beta_3 \text{segregation}_i$$

Estimate this model in R

```r
budgets_m1 <- lm(exp_police_pc ~ rev_prop_tax_pc + violent.crime.high + segregation.bw.high,
    data = budgets)

tidy(budgets_m1)
```

```
## # A tibble: 4 x 5
##   term                    estimate std.error statistic  p.value
##   <chr>                      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)               116.      10.6       11.0  2.17e-25
## 2 rev_prop_tax_pc             0.168    0.00744   22.6  3.20e-78
## 3 violent.crime.highTRUE      6.60    11.0        0.598 5.50e- 1
## 4 segregation.bw.highTRUE     6.85    11.2        0.613 5.40e- 1
```
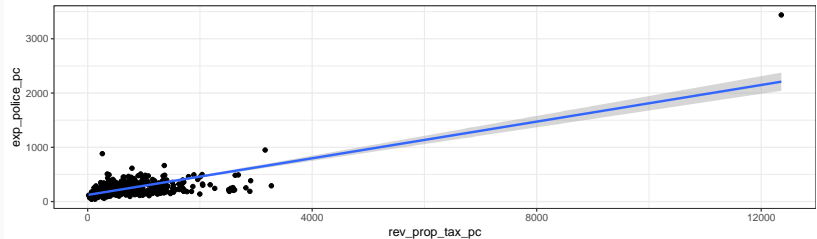
1. A linear model is a reasonable approximation of the data generating process
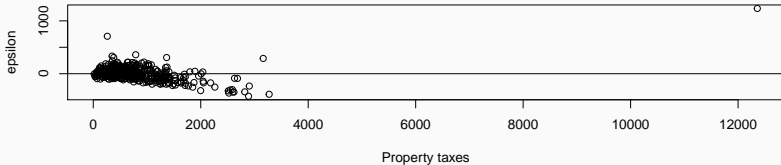
Q: Is police spending a linear function of property taxes?

```
ggplot(budgets, aes(x = rev_prop_tax_pc, y = exp_police_pc)) + geom_point() + geom_smooth(method = "lm")
```

2. Constant error variance (Homoskedasticity): $V(\varepsilon|X) = V(\varepsilon)$

```
plot(budgets_m1$model$rev_prop_tax_pc, budgets_m1$residuals, xlab = "Property taxes",
    ylab = "epsilon")
abline(0, 0)
```

Revise the model!

Model assumptions

1. A linear model is a reasonable approximation of the data generating process. *Not really! Let's try a logarithm of these rate per capita variables*

1. A linear model is a reasonable approximation of the data generating process. *Not really! Let's try a logarithm of these rate per capita variables*

Q: Is police spending a linear function of property taxes after log transformation?

```
ggplot(budgets, aes(x = log(rev_prop_tax_pc), y = log(exp_police_pc))) + geom_point() +
    geom_smooth(method = "lm")
```
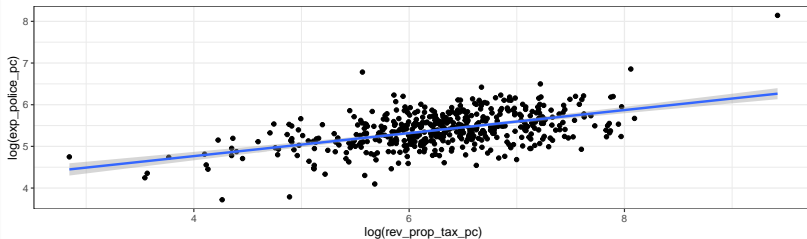
Log transformations
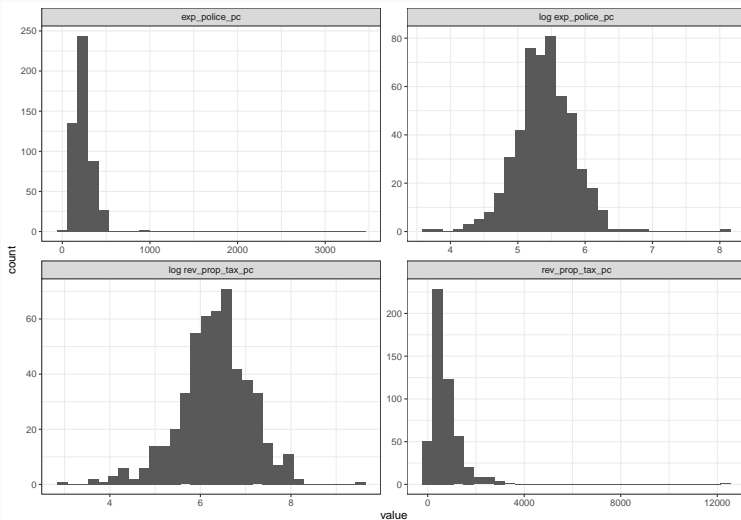
Fit a new model

```
budgets_m2 <- lm(log(exp_police_pc) ~ log(rev_prop_tax_pc) + violent.crime.high +
    segregation.bw.high, data = budgets)

tidy(budgets_m2)
```

```
## # A tibble: 4 x 5
##   term                    estimate std.error statistic   p.value
##   <chr>                      <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)                 3.66     0.134     27.3  4.89e-101
## 2 log(rev_prop_tax_pc)        0.270    0.0213    12.7  3.87e- 32
## 3 violent.crime.highTRUE      0.0187   0.0334     0.559 5.76e-  1
## 4 segregation.bw.highTRUE     0.0661   0.0337     1.96  5.07e-  2
```
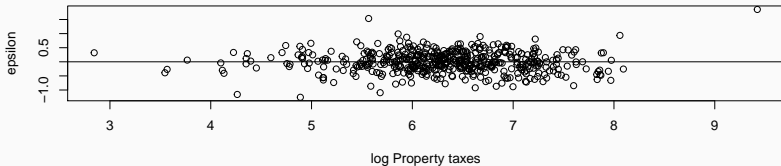
2. Constant error variance (Homoskedasticity): $V(\varepsilon|X) = V(\varepsilon)$

```
plot(budgets_m2$model$`log(rev_prop_tax_pc)`, budgets_m2$residuals, xlab = "log Property taxes",
    ylab = "epsilon")
abline(0, 0)
```

Interpreting the model (descriptive)

```r
tidy(budgets_m2)
```

```
## # A tibble: 4 x 5
##   term                  estimate std.error statistic  p.value
##   <chr>                    <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)              3.66      0.134     27.3   4.89e-101
## 2 log(rev_prop_tax_pc)     0.270     0.0213    12.7   3.87e- 32
## 3 violent.crime.highTRUE   0.0187    0.0334     0.559 5.76e-  1
## 4 segregation.bw.highTRUE  0.0661    0.0337     1.96  5.07e-  2
```

We predict that for counties in this sample with similar levels of property
taxes, high segregation counties generally spend more on police than do low
segregation counties: 95% CI: $e^{\hat{\beta} \pm z_{\alpha/2} \times SE_\beta} =$ [-0.0038384, 14.1282868]

```r
## create possible values for prediction
rev_prop_tax_pc <- seq(from = quantile(budgets$rev_prop_tax_pc, 0.05), to = quantile(budgets$rev_prop_tax_
    0.95), length.out = 1000)

violent.crime.high <- c(T, F)
segregation.bw.high <- c(T, F)

## make all combinations possible

new_dat <- expand_grid(rev_prop_tax_pc, violent.crime.high, segregation.bw.high)
## calculate predictions
preds <- as.data.frame(predict(budgets_m2, newdata = new_dat, interval = "confidence"))
new_dat <- new_dat %>% mutate(yhat = preds$fit, yhat_lwr = preds$lwr, yhat_upr = preds$upr)
```
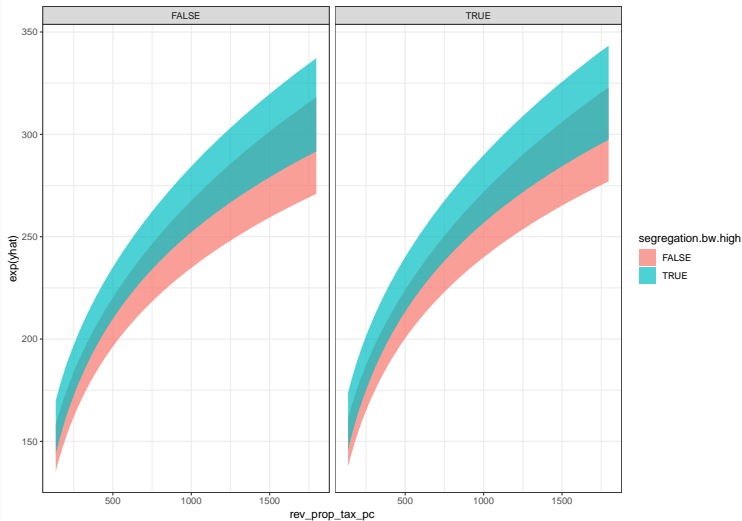
Interpreting the model using prediction (simulation). Police spending by property tax revenue, segregation, and violent crime levels (high = T,F)

Interpretation in this context

- These results are not the causal effect of segregation or of property taxes! We've made no effort to address unmeasured confounders of property taxes, segregation, and police spending (there are many!)
- They do provide information on the direction, magnitude, and precision of relationships in the observed data
- In these cases, higher property taxes tend to mean higher police spending. Levels of spending tend to be higher in segregated counties.
- Confidence intervals provide us with information about the precision of the estimated relationship
- Don't over-interpret β: focus on direction (+/-), magnitude (big, little), precision (noisy, consistent)
- Think about how predictors move together in the real data to constrain your predictions and make them more reasonable. Predictors are correlated!

- HW 9 is in the mother-wage-penalty folder, along with the data you need for the exercise
- See the IPV example in lecture 7 for more on fixed effects models
- Please complete the course eval if you haven't
- HW 9 is due 12/18 by 10AM.
- Thanks for a great semester!