Uncertainty, 2

Frank Edwards

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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

The linear regression model

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

The linear regression model

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

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

The linear regression model

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

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

The linear regression model

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

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 linear regression model

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

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 $E(y_i) = \beta_0 + \beta_1 x_i$ provides a prediction for the values of y_i based on the values of x_i.

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

The linear regression model and prediction

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

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

The linear regression model and prediction

Remember, that we put a $\hat{\textit{ha}}\textit{t}$ on variables to indicate that they are estimated from the data, or predicted.

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

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

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

The linear regression model and prediction

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

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

A regression line predicts values of Y, by estimating $\hat{Y}$ 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

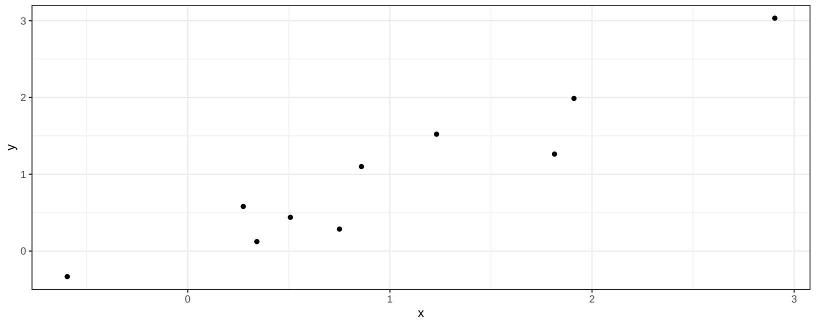
```
## # A tibble: 10 x 2
##         x      y
##     <dbl>  <dbl>
##  1  0.342  0.123
##  2  0.508  0.439
##  3  1.23   1.52
##  4  1.91   1.99
##  5  1.81   1.26
##  6  0.859  1.10
##  7 -0.596 -0.333
##  8  0.275  0.580
##  9  2.90   3.03
## 10  0.751  0.286
```

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

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

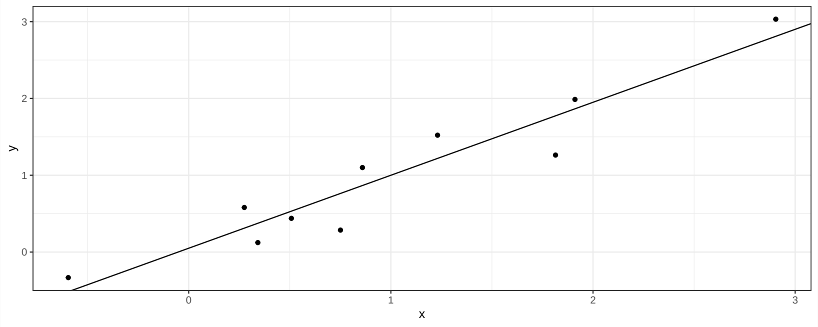
Understanding the regression line

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



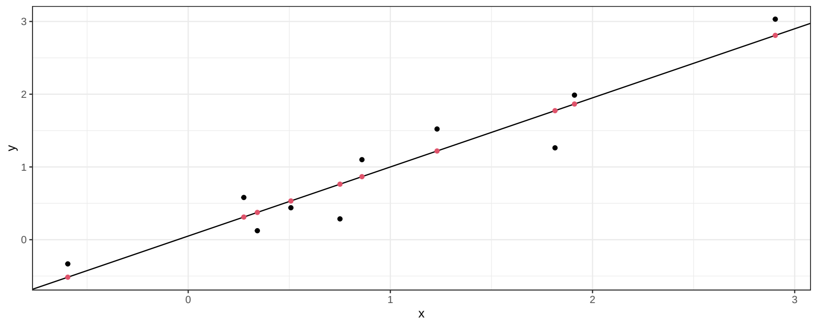
Understanding the regression line: adding the fit

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



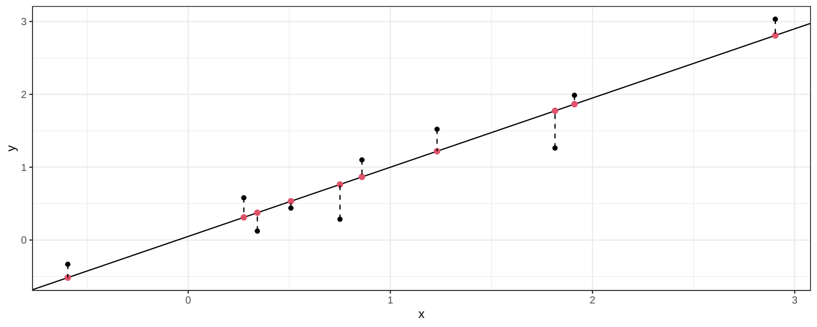
Understanding the regression line: adding $\hat{y}$

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



Understanding the regression line: adding ε

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



Goal of linear regression

1. Estimate causal relationships between some predictor variable x and outcome variable y

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

Assumptions of a linear regression model (for causal estimates)

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

1. The linear model approximates the data generating process

Assumptions of a linear regression model (for causal estimates)

For estimates of β to be unbiased and 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

Assumptions of a linear regression model (for causal estimates)

For estimates of β to be unbiased and 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
3. Linear independence of predictors

Assumptions of a linear regression model (for causal estimates)

For estimates of β to be unbiased and 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
3. Linear independence of predictors
4. Constant error variance (Homoskedasticity): $V(\varepsilon|X) = V(\varepsilon)$

Assumptions of a linear regression model (for prediction)

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

Assumptions of a linear regression model (for prediction)

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): $V(\varepsilon|X) = V(\varepsilon)$

Guidance

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*.

Guidance

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:

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

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

Ways to express an OLS model

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

The Mark of a Criminal Record

```
### read and format Pager data
cr<-read_csv("https://raw.githubusercontent.com/f-edwards/intro_stats/master/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
```

The sample average treatment effect and its standard error

```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))

SATE
```

```
## # A tibble: 2 x 5
##   race    xbar     se    SATE SATE_se
##   <chr>  <dbl>  <dbl>   <dbl>   <dbl>
## 1 Black 0.0508 0.0157 -0.0899   0.0293
## 2 White 0.167  0.0305 -0.173    0.0494
```

Thinking through the OLS assumptions

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

1. The linear model approximates the data generating process: *Questionable under binary outcome, but ok*

For estimates of β to be unbiased and 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*

Thinking through the OLS assumptions

For estimates of β to be unbiased and 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*

Thinking through the OLS assumptions

For estimates of β to be unbiased and 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*

Standard errors of β

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.

Standard errors of β

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:

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

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:

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

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

If the sampling distribution for β is defined as:

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

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:

$H_0 : \beta \sim N(0, SE^2_{\beta})$

Using OLS to estimate the SATE

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


summary(cr_ols)
```

```
##
## Call:
## lm(formula = callback ~ race * crimrec, data = cr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34000 -0.16667 -0.14070 -0.05076  0.94924
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.14070    0.02529   5.565 3.76e-08 ***
## raceWhite              0.19930    0.03857   5.167 3.11e-07 ***
## crimrecTRUE           -0.08994    0.03585  -2.509   0.0123 *
## raceWhite:crimrecTRUE -0.08339    0.05460  -1.527   0.1272
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3567 on 692 degrees of freedom
## Multiple R-squared:  0.07638,    Adjusted R-squared:  0.07238
## F-statistic: 19.08 on 3 and 692 DF,  p-value: 6.781e-12
```

Predicting outcomes from our OLS model

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

```
## # A tibble: 4 x 2
##   race  crimrec
##   <chr> <lgl>
## 1 Black FALSE
## 2 Black TRUE
## 3 White FALSE
## 4 White TRUE
```

Predicting outcomes from our OLS model

```
### generate predictions


yhat<-predict(cr_ols,
              newdata = pred_dat,
              interval = "confidence")

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

head(cr_ols_yhat)
```

```
##          fit          lwr       upr
## 1 0.14070352 0.0910575899 0.1903494
## 2 0.05076142 0.0008641203 0.1006587
## 3 0.34000000 0.2828173156 0.3971827
## 4 0.16666667 0.1094839823 0.2238494
```

```
### append these predictions to the prediction data

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

Check results

```
(pred_dat)
```

```
## # A tibble: 4 x 5
##   race  crimrec   yhat yhat_lwr yhat_upr
##   <chr> <lgl>    <dbl>    <dbl>    <dbl>
## 1 Black FALSE   0.141  0.0911      0.190
## 2 Black TRUE    0.0508 0.000864    0.101
## 3 White FALSE   0.340  0.283       0.397
## 4 White TRUE    0.167  0.109       0.224
```

```
(treatment)
```

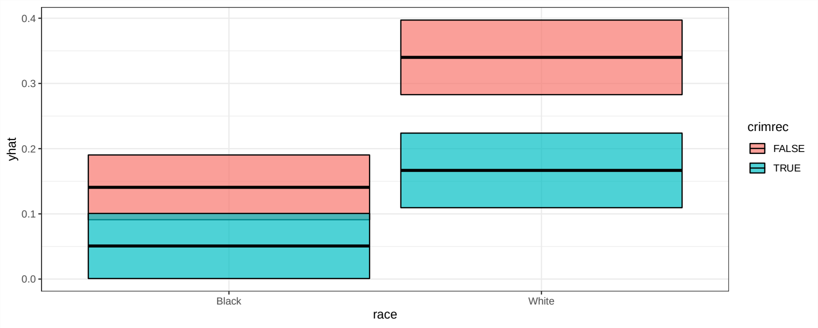
```
## # A tibble: 2 x 3
##   race    xbar     se
##   <chr>  <dbl>  <dbl>
## 1 Black 0.0508 0.0157
## 2 White 0.167  0.0305
```

```
(control)
```

```
## # A tibble: 2 x 3
##   race   xbar     se
##   <chr> <dbl>  <dbl>
## 1 Black 0.141 0.0247
## 2 White 0.34  0.0388
```

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)
```



Methods using OLS to obtain causal estimates

1. Randomized controlled trial

Methods using OLS to obtain causal estimates

1. Randomized controlled trial
2. “Natural experiments”: Difference in differences, other matched-group methods (propensity scores, etc)

Methods using OLS to obtain causal estimates

1. Randomized controlled trial
2. “Natural experiments”: Difference in differences, other matched-group methods (propensity scores, etc)
3. Instrumental variables

Methods using OLS to obtain causal estimates

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

```r
budgets<-read.csv("https://raw.githubusercontent.com/f-edwards/intro_stats/master/data/police_spending.csv")
glimpse(budgets)
```

```
## Rows: 501
## Columns: 7
## $ fips                <int> 1001, 1021, 1033, 1061, 1081, 1083, 1089, 1097, 11~
## $ exp_police_pc       <dbl> 169.24, 170.18, 192.94, 164.06, 186.56, 163.66, 21~
## $ officers_pc         <dbl> 83.36, 32.19, 31.90, 29.89, 51.58, 34.35, 19.80, 1~
## $ rev_prop_tax_pc     <dbl> 158.17, 242.31, 256.04, 260.15, 330.72, 169.51, 36~
## $ violent.crime.high  <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, FALSE, ~
## $ segregation.bw.high <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, TRU~
## $ segregation.lw.high <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, TRUE, TRUE, FALSE~
```

Develop a theoretical model to predict police department budgets

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

Develop a theoretical model to predict police department budgets

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

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)
```

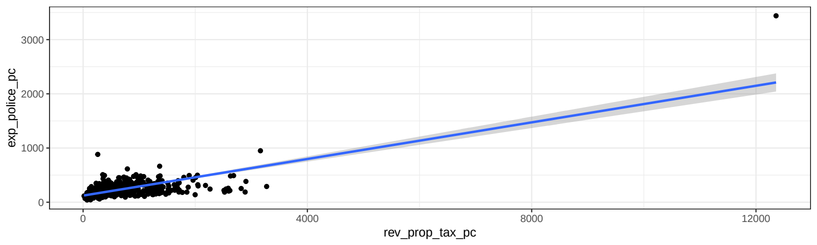
```
## # A tibble: 4 x 5
##   term                    estimate std.error statistic  p.value
##   <chr>                      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)              116.      10.6        11.0   2.17e-25
## 2 rev_prop_tax_pc            0.168    0.00744    22.6   3.20e-78
## 3 violent.crime.highTRUE     6.60    11.0         0.598 5.50e- 1
## 4 segregation.bw.highTRUE    6.85    11.2         0.613 5.40e- 1
```

Check the model assumptions

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")
```



Check the model assumptions

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!

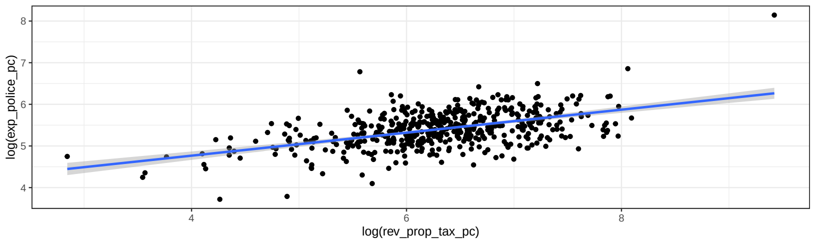
1. A linear model is a reasonable approximation of the data generating process. *Not really! Let's try a logarithm of these rate per capita variables*

Model assumptions

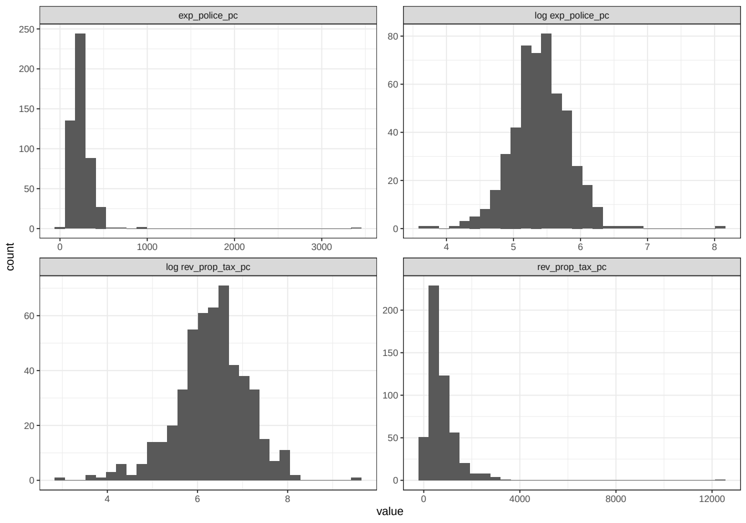
1. A linear model is a reasonable approximation of the data generating process. *Not really! Let's try a logarithm of these rate per capita variables*

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

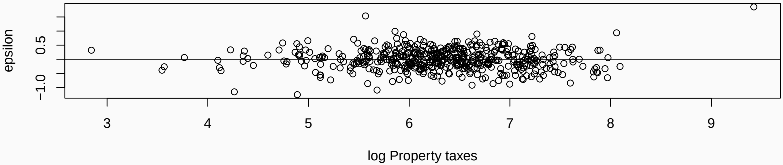
```
budgets_m2<-lm(log(exp_police_pc) ~
                 log(rev_prop_tax_pc) +
                 violent.crime.high +
                 segregation.bw.high,
               data = budgets)

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, 0.14]

Interpreting the model using prediction (simulation)

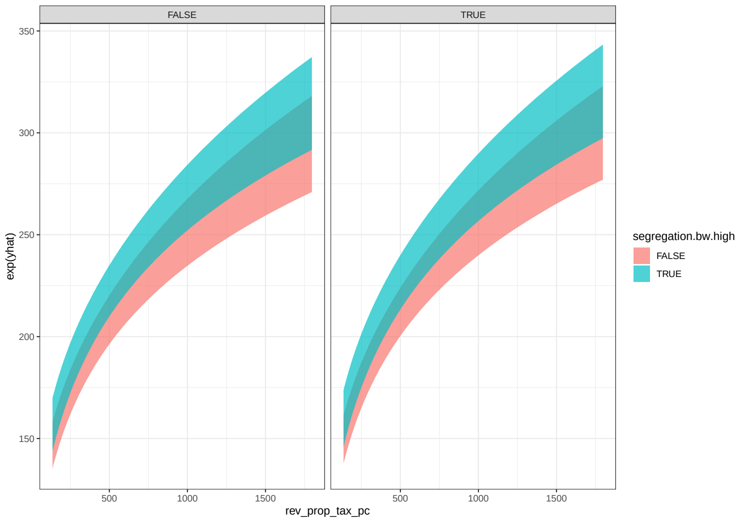
```r
## create possible values for prediction
rev_prop_tax_pc<-seq(from = quantile(budgets$rev_prop_tax_pc, 0.05),
              to = quantile(budgets$rev_prop_tax_pc, 0.95),
              length.out = 1000)

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!)

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

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.

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.

Interpretation in this context

- Confidence intervals provide us with information about the precision of the estimated relationship

Interpretation in this context

- 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)

Interpretation in this context

- 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 11 is in the mother-wage-penalty folder, along with the data you need for the exercise
- Please complete the course eval if you haven't