Session 7: Basic Statistical Models

An introduction to:

- Simple Linear Regression,
- Multiple Regression, and
- Logistic Regression

Prediction

Statistics is not only about data description and statistical testing. Another very important aspect is prediction.

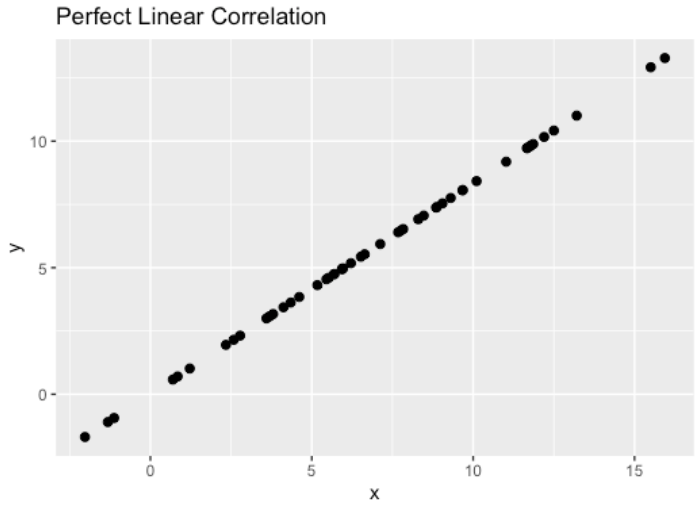
Regression analysis aims to predict a continuous dependent variable y with the help of at least one independent variable x.

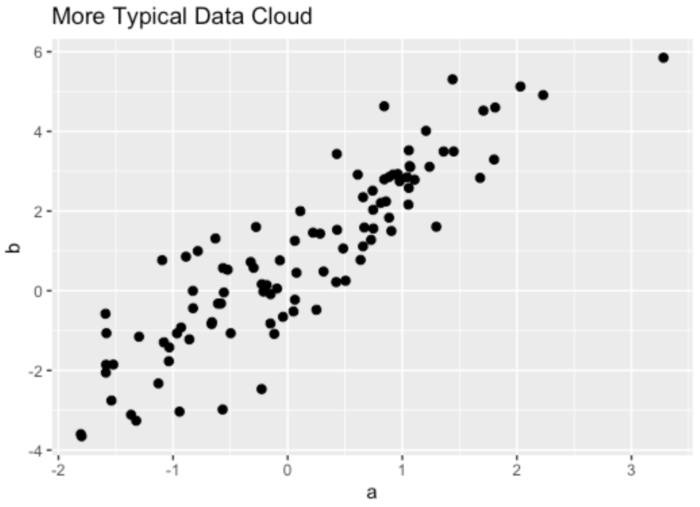
For example:

- How does education x affect future earnings y?
- Does police presence x impact the crime rate y?

Correlation

The precision of a prediction depends on the correlation between *x* and *y*. **But remember, correlation does not imply causation.**



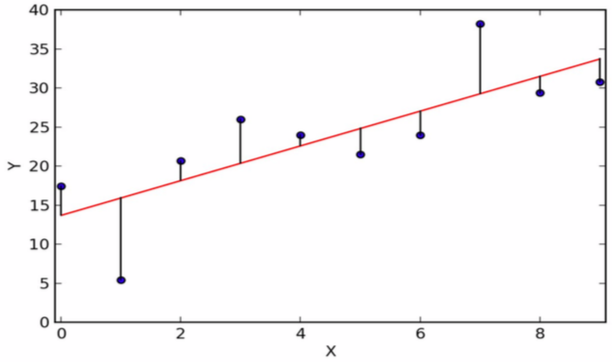


Ordinary Least Squares

How do we use this data cloud to predict *y* from *x*?

We find the best line that summarizes the relationship, called the regression line.

1. Draw a hypothetical line.
2. Measure the distance between each observation and the line. This is called the "residual".
3. Square each residual (to cancel out negatives and punish outliers)
4. Add all of the squared residuals together.
5. Draw another hypothetical line, and repeat the process.
6. Ultimately, choose the line with the lowest total sum of the squared errors.



Interpreting the Regression Line

A straight line is defined by its slope and intercept on the y-axis: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$

- ❖ $\hat{y}_i$ is the estimated value of *yi* (= *y* for observation *i*)
- ❖ $\hat{\beta}_0$ is the "intercept," or value of *y* when *x* is 0
- ❖ $\hat{\beta}_1$ Is the slope, or the change in *y* that results from a one-unit change in *x*

Imagine $\hat{\beta}_0 = 10$, $\hat{\beta}_1 = 3$, and $x_i = 5$? What would $\hat{y}_i$ be?

** The "hats" denote that these are estimates of the regression line*

From Regression Line to Regression Model

To make the regression line become a regression model, we use the true observation *yi,* and add the prediction error $\hat{e}_i$

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i + \hat{e}_i$$

The other coefficients remain the same.

$\hat{e}_i$ is the *residual* (or estimated error term, i.e. the difference between the predicted value of *y* and the actual *y* value). It contains the effects of all the variables that influence both *y* and *x*, but cannot be observed (omitted variables), as well as random measurement errors.

Creating a Linear Model in R

The syntax to create a linear model in R is simple:

```
model <- lm(y ~ x, data = data)
```

From here, we can view our results by calling:

```
summary(model)
```

Practice time!

Let's estimate a simple linear model to predict math test scores based on the number of teachers at a school:

```
library(AER)

data("CASchools")

model <- lm(math ~ teachers, data = CASchools)

summary(model)
```

Multiple Regression

Up until now, we've only looked at the effect of one *x* variable on *y*. But most of the time there are lots of things that are related to an outcome.

With multiple regression, we can estimate a separate coefficient for each variable.

$$y_i = \hat{\beta_0} + \hat{\beta_1} \cdot x_i + \hat{\beta_2} \cdot x_i + \hat{\beta_3} \cdot x_i + \hat{e_i}$$

Now each beta coefficient represents a partial association between the outcome and the *x* variable, controlling for the other explanatory variables.

The syntax in R is almost the same, too:

```
model <- lm(y ~ x + z, data = data)
```

Practice time!

Let’s return to our student achievement example. What could be driving the negative relationship between student test scores and the number of teachers in a school? What confounders might there be?

Some variables to try:

`income`

`students`

`lunch` (a measure of low income)

`english` (percentage of English learners)

`expenditure`

APPENDIX

Here are some extra resources on regression assumptions, which you really ought to keep in mind when conducting regression analysis ;-)

Regression Assumptions

In order for the least squares estimator to be the “Best Linear Unbiased Estimator” (BLUE) and for our linear regression results to be reliable, several **assumptions** have to hold:

- Linearity: The relationship between x and y must be linear.
- Weak exogeneity: The independent variable x is not random, but deterministic, and must be independent of the error term.
- Independence of errors: The residuals must be uncorrelated.
- The expectation of the errors is 0: The expected value of the residuals is 0, and should randomly vary around 0.
- Homoscedasticity: The variance of the prediction error must be constant.
- Normality of errors: The residuals should be normally distributed (optional assumption that is not needed for OLS, but is relevant for significance testing).

Checking the Assumptions

A well-established way to check for violations of the assumptions is to use a residual plot. Residual plots show the predicted values for *y* on the *x*-axis, and the estimated residuals on the *y*-axis.

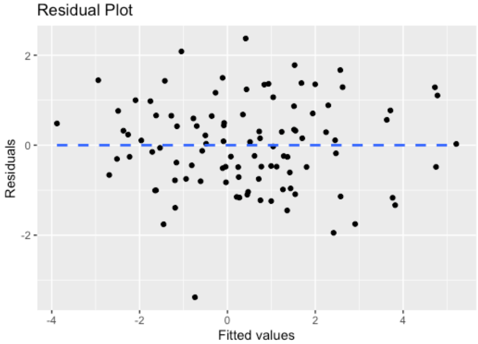
The simplest (though not prettiest) way to do this is with: **`plot(model)`**

In `ggplot2`:
```
ggplot(data, aes(x = predict(model),
                 y = residuals(model))) +
                     geom_point() +
                     stat_smooth(method = "lm",
               se = FALSE, linetype = "dashed") +
                     labs(title = "Residual plot",
               x = "Fitted values", y = "Residuals")
```

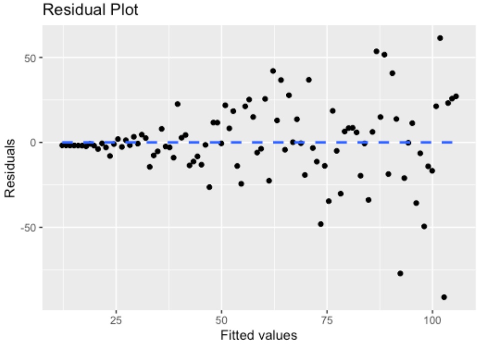
Checking the Assumptions

You want to see a random cloud of residuals scattered around 0.

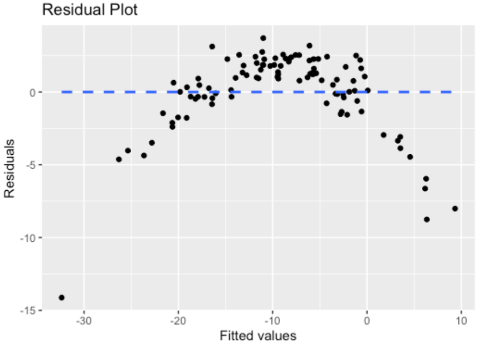
Like this:



Not like this:



This is also bad:



Linear Regression with a binary variable

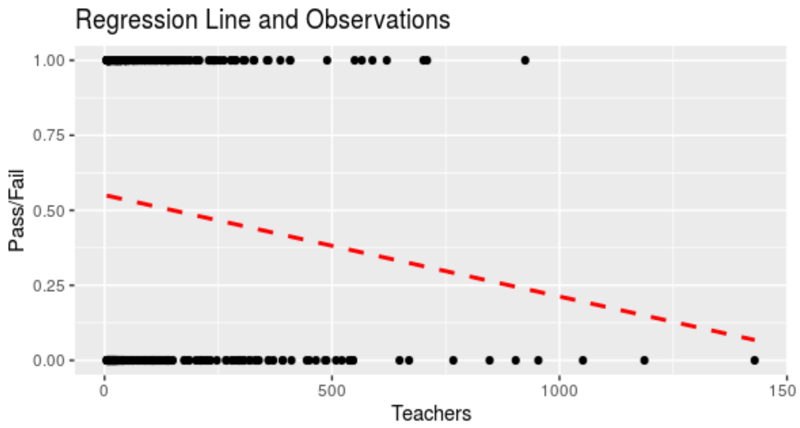
How do results change when the dependent variable is no longer continuous , but categorical e.g. binary? Let's predict students' test scores based on the number of teachers at a school again. This time we recode the scores to be pass (1) or fail (0):

Code snippet:

```
CASchools <- CASchools %>%
mutate(score = (math + read) / 2,
score2 = ifelse(score >= mean(score), 1, 0)

model <- lm(score2 ~ teachers, data =
CASchools)

summary(model3)
```



Linear Regression with a binary variable

- Predicting a categorical variable for an observation == assigning the observation to a class (classification)
- Hence we predict the probability of each of the classes of the categorical variable
- Linear Regression is not capable of predicting probability
- The linear regression model represents these probabilities as: p(X)=β0 + β1X
- Predicted values could fall out of range of possible values [0 – 1]
- To avoid this problem, use the logistic function to model p(X) that gives outputs between 0 and 1 for all values of X (see next slide):

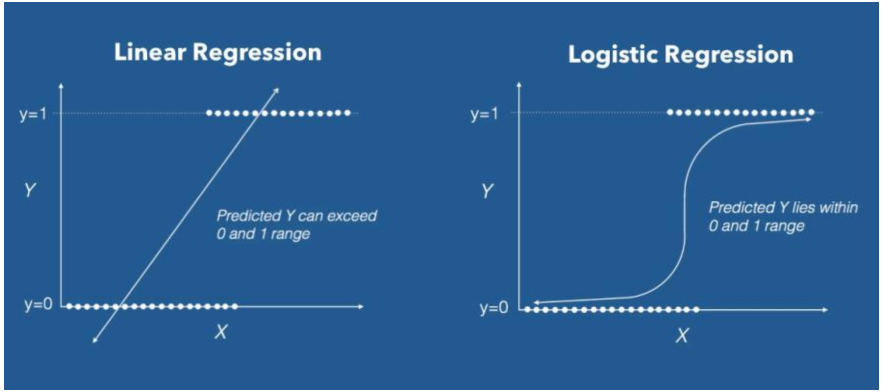
$$f(x) = \frac{1}{1+e^{-x}} \Rightarrow p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \Rightarrow \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

(Odds)

Logistic Regression

…similar to linear regression, except that the dependent variable is categorical and not continuous. For instance, we predict whether students pass or fail.

- Instead of fitting a line to the data, logistic regression fits an S-shaped curve produced by the logistic function



Logistic Function:

$$f(x) = \frac{1}{1+e^{-x}}$$

- The curve goes from 0 to 1; It tells you the probability of outcome Y (e.g. pass) based on X (e.g. teachers)
- Just like linear regression, logistic regression can work with continuous and discrete *independent* variables.

Interpreting Coefficients

- Remember our equation: $$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$
- When we take the log odds of both sides we get: $$\underbrace{log(\frac{p(X)}{1-p(X)})}_{\text{log(odds) or logit}} = \beta_0 + \beta_1 X$$
- Coefficients are presented in terms of log(odds)

```
Call:
glm(formula = score2 ~ teachers + income2, family = binomial(),
    data = CASchools)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2341  -0.8201   0.4153   1.0482   2.1697

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.3853612  0.2640255  -5.247 1.55e-07 ***
teachers      -0.0028084  0.0007575  -3.708 0.000209 ***
income2Middle  1.7280964  0.2992350   5.775 7.69e-09 ***
income2High    3.8106453  0.4102626   9.288  < 2e-16 ***
```

Holding other variables constant:

- For every additional teacher, the log(odds of passing) decreases by 0.002
- The log-odds of passing are 1.728 higher when the student is from a middle-income district compared to when he is from a low-income district
- The log-odds of passing are 3.810 higher when the student is from a high-income district compared to when he is from a low-income district

Interpreting Coefficients

- Differences in the log-odds are difficult to conceptualize. In general, do not interpret them.
- Use odds-ratio instead
- Convert *log-odds* differences to *odds-ratios* by taking the exponential of the coefficients

```
> #odds ratio
> model4_OR <- exp(coef(model4))
> model4_OR
(Intercept)     teachers income2Middle   income2High
  0.2502334    0.9971955     5.6299265    45.1795838
```

Holding other variables constant:

- An additional teacher decreases the student's odd of passing by 1%.
- The odds of passing are higher for students from a middle income district compared to those from a low income district. The odds are about 462% higher for students from a middle income district than those from a low income district
- The odds of passing are higher for students from a high income district compared to those from a low income district. The odds are about 4417% higher for students from a high income district than those from a low income district.

Building the Model

- You fit the model using the *glm()* function

The syntax in R is almost the same as the linear model, too:

```
model <- glm(y ~ x + z, data = data, family = binomial())
```

Practice:

Let's return to our student achievement example.

- Use both math and english to determine students scores (i.e. score = (math + english) / 2
- Generate a new categorical variable from the score with class *Pass* or *Fail. Your threshold should be the mean of the scores.*
- Generate a categorical variable from the *income* variable with classes *Low, Middle,* and *High.*
- Model the relationship between the score (binary dependent variable), the number of teachers (continuous independent variable) and the average district income (categorical independent variable).