Lecture 6: Linear regression

Criminology 250

Prof Maria Cuellar

University of Pennslyvania

What is linear regression?

The correlation between two variables tells us whether the linear association between them is strong. But it does not tell us *what the line is*.

A linear model gives an equation of a straight line through the data. This model can predict the value y for any value x, which can be within the sample or not.

Of course, no line will go through all the points, but a linear model can summarize the general pattern with only a couple of parameters.

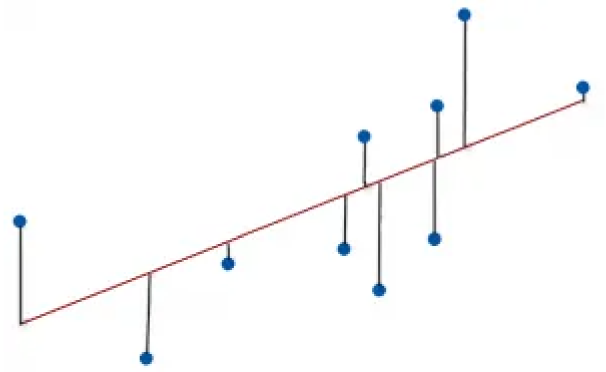
Like all models of the real world, the line will be wrong in the sense that it can't match reality *exactly*, but it can help us understand how the variables are associated.

Predicted values

The estimate made from a model is the *predicted value* and we write it as $\hat{y}$ (called *y-hat*) to distinguish it from the observed value, *y*.

The difference between the observed value and its associated predicted value is called the *residual*. The residual value tells us how far off the model's prediction is at that point.

$$Residual = Observed\ value - Predicted\ value.$$



Residuals are the distance between the observed value and the fitted value.

The linear model

A straight line can be written as

$$y = mx + b.$$

We'll use this form for our linear model, but in statistics we use slightly different notation:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x.$$

We write $\hat{y}$ to emphasize that the points that satisfy this equation are just our *predicted* values, not the actual data values, which scatter around the line.

We write b_0 and b_1 for the slope and intercept of the line. The estimated $\hat{b}$'s are called the *coefficients* of the linear model, b_1 is the slope which tells how rapidly $\hat{y}$ changes with respect to x, and b_0 is the intercept, which tells where the line intercepts the y-axis.

Interpretation

Suppose a model is:

$$\widehat{Fat} = 8.4 + 0.91 Protein.$$

This means the slope 0.91 says that an item with one more gram of protein can be expected, on average, to have 0,91 more grams of fat. For the intercept, even without protein, an item would have, on average, 8.4 grams of fat. Often the intercept won't be meaningful, but it helps us position the model vertically.

This interpretation works for continuous (quantitative) predictor and outcome, but it's not the same if either changes. More on this later.

Warning

This is where "correlation does not imply causation" becomes very important. It is not the same to say

"If you increase protein by one gram, fat will increase by one gram," as

"For a higher protein, by one gram, fat is higher by one gram, on average."

(The second one is correct!) Why are these different?

Recall: Assumptions of linear regression

1. Homoscedasticity: Homogeneity of variance. The size of the error in our prediction doesn’t change significantly across the values of the independent variable. i.e., The variance of residuals is the same for any value of x.

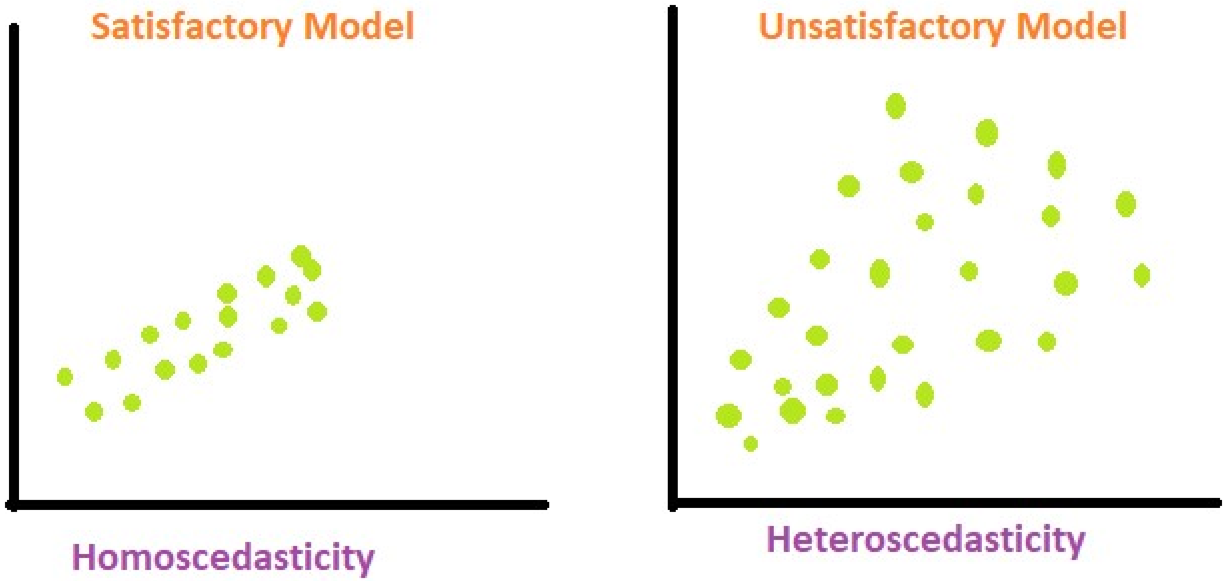
2. Independence between observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.

3. Normality: The data follows a normal distribution. i.e., For any fixed value of x, y is normally distributed.

4. Linearity: The relationship between the independent and dependent variable is linear: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor). i.e., The relationship between X and the mean of Y is linear.

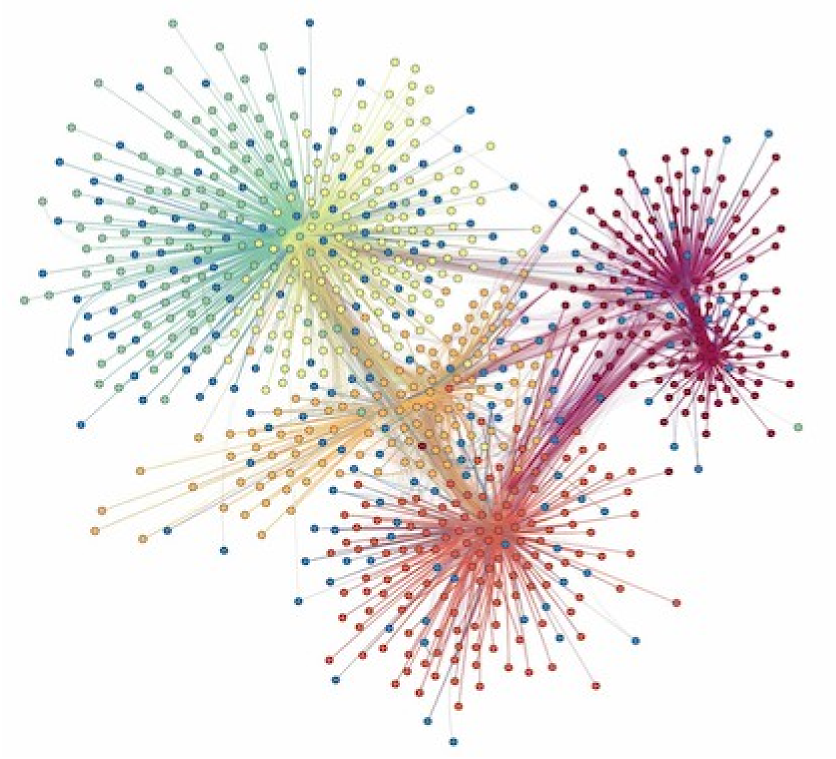
Visualizing assumption 1: Homoscedasticity

The assumption is that the data is homoscedastic.



Visualizing assumption 2: Independence of observations

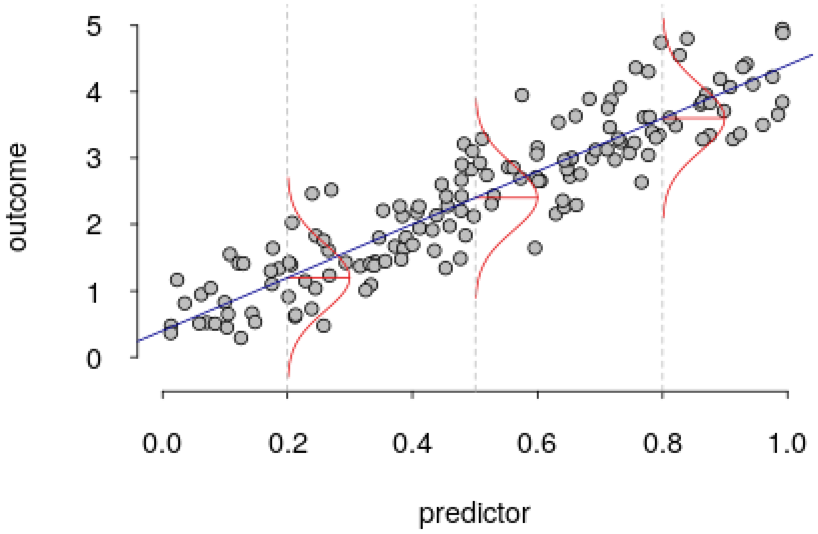
We do not want connections like the ones in this social network:



Visualizing assumption 3: Normality

Visualizing assumption 4: Linearity

Actually, this is linearity, homoscedasticity, and normality.



Diagnosic plots

Diagnosic plots are an amazing tool for checking whether your regression model is fitting well and satisfying assumptions.

We will study four of them that come with the `lm` function in R:

- Goodness of fit:
  - R-squared: Coefficient of determination
  - Residuals vs. fitted
- Assumption 1: Homoscedasticity
  - Scale-location
- Assumption 2: Independence
  - Can't test visually
- Assumption 3: Normality
  - Normal Q-Q plot
  - Residuals vs. leverage
- Assumption 4: Linearity
  - No official plot, but could use a scatterplot.

Goodness of fit: Coefficient of determination

How can we tell if a linear model is fitting properly?

The coefficient of determination or R^2 is the variation accounted for by the model. $R^2 = 1$ means your model perfectly predicts the data, i.e., that all of the variance in the data is in the model. $R^2 = 0$ means your model is not a good fit for the data.

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}.$$

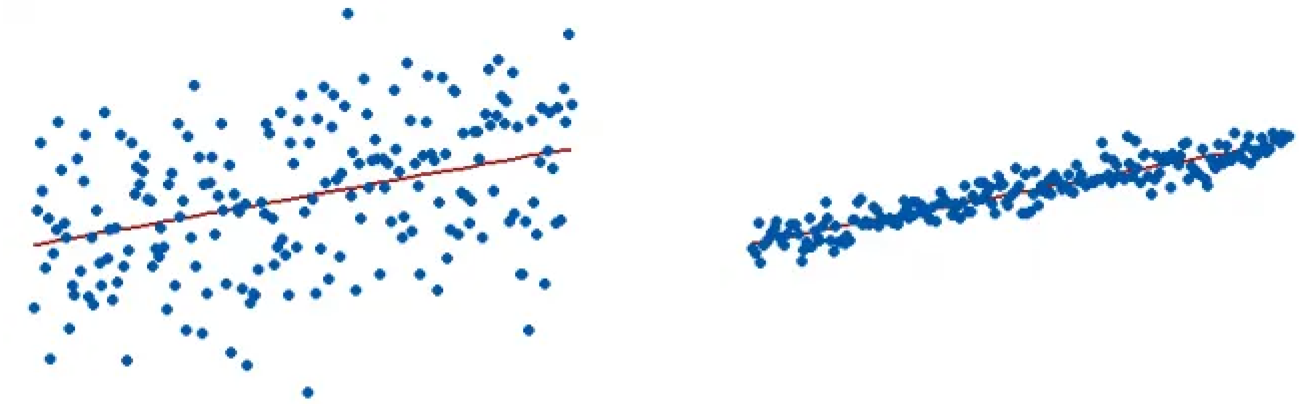
It is read as a percentage (i.e., has a value from 0 to 100%).

Goodness of fit: Coefficient of determination

The R^2 on the left is 15% and the R^2 on the right is 85%.

When a regression model accounts for more of the variance, the data points are closer to the regression line. In practice, you'll never see a regression model with an R2 of 100%.

Regression models with low R-squared values can be perfectly good models. Some fields of study have an inherently greater amount of unexplainable variation. In these areas, R^2 values are bound to be lower.



Scale-location plot: A way to test assumption 1 (homoscedasticity)

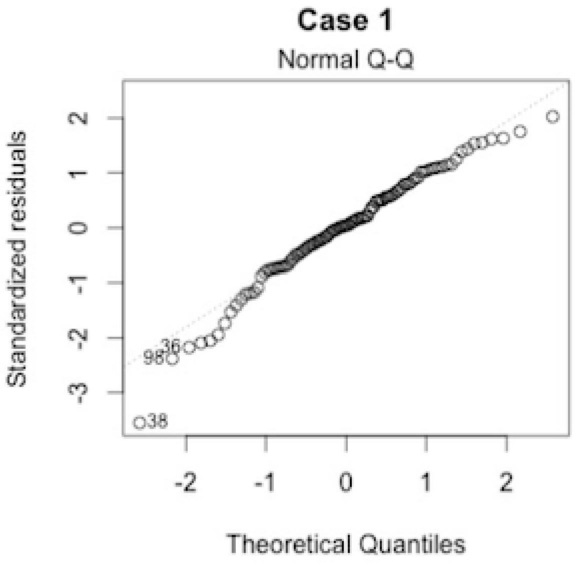
It’s also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It’s good if you see a horizontal line with equally (randomly) spread points.

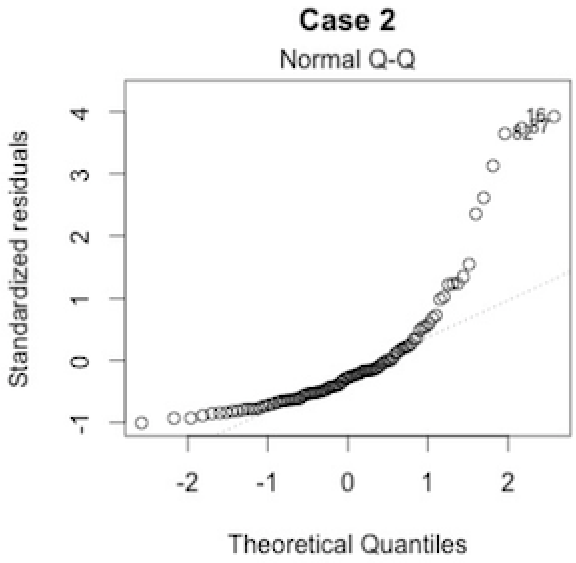
Residuals vs. fitted plot: A more informative version of R^2

We want to know how well the model fits, so we can ask instead what the model missed: the residuals. A scatterplot of the residuals vs. the x-values should be the most boring scatterplot you've ever seen: It shouldn't have any interesting features, like a direction, shape, or outliers (left figure).

Normal Q-Q plot: A way to test assumption 3 (normality)

This plot shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It’s good if residuals are lined well on the straight dashed line. What do you think? They are probably never a perfect straight line, and this will be your call. Case 2 below definitely concerns me. I would not be concerned by Case 1 too much, although an observation numbered as 38 looks a little off.





Residuals vs. leverage plot: A way to test for points of high leverage

This plot helps us to find influential cases, if there are any. Not all outliers are influential. Unlike the other plots, here patterns are not relevant. We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential for the model.

Look for cases outside of a dashed line, Cook’s distance. When cases are outside of the Cook’s distance (meaning they have high Cook’s distance scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.

Residuals vs. leverage plot

Case 1 is the typical look when there is no influential case, or cases. You can barely see Cook's distance lines (a red dashed line) because all cases are well inside of the Cook's distance lines. In Case 2, a case is far beyond the Cook's distance lines (the other residuals appear clustered on the left because the second plot is scaled to show larger area than the first plot). The plot identified the influential observation as #49. If I exclude the 49th case from the analysis, the slope coefficient changes from 2.14 to 2.68 and R^2 from .757 to .851. Pretty big impact!

Other tools to improve model fit

Your current model might not be the best way to understand your data. In that case, you may want to go back to the beginning. Is it really a linear relationship between the predictors and the outcome?

There are other tools we have not covered yet, which might make the model fit the data better:

- You may want to include a quadratic term, for example.
- A log transformation may better represent the phenomena that you'd like to model.
- Or, is there any important variable that you left out from your model? Other variables you didn't include (e.g., age or gender) may play an important role in your model and data.
- Or, maybe, your data were systematically biased when collecting data. You may want to redesign data collection methods.

(Source: https://data.library.virginia.edu/diagnostic-plots/)

Example: Simulated height and weight data

```r
x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)
out <- lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3002 -1.6629  0.0412  1.8944  3.9775
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -38.45509    8.04901  -4.778  0.00139 **
## x             0.67461    0.05191  12.997 1.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
##
## Residual standard error: 3.253 on 8 degrees of freedom
## Multiple R-squared:  0.9548,    Adjusted R-squared:  0
## F-statistic: 168.9 on 1 and 8 DF,  p-value: 1.164e-06
```

speed.c = scale(cars$speed, center=TRUE, scale=FALSE) mod1 = lm(formula = dist ~ speed.c, data = cars)

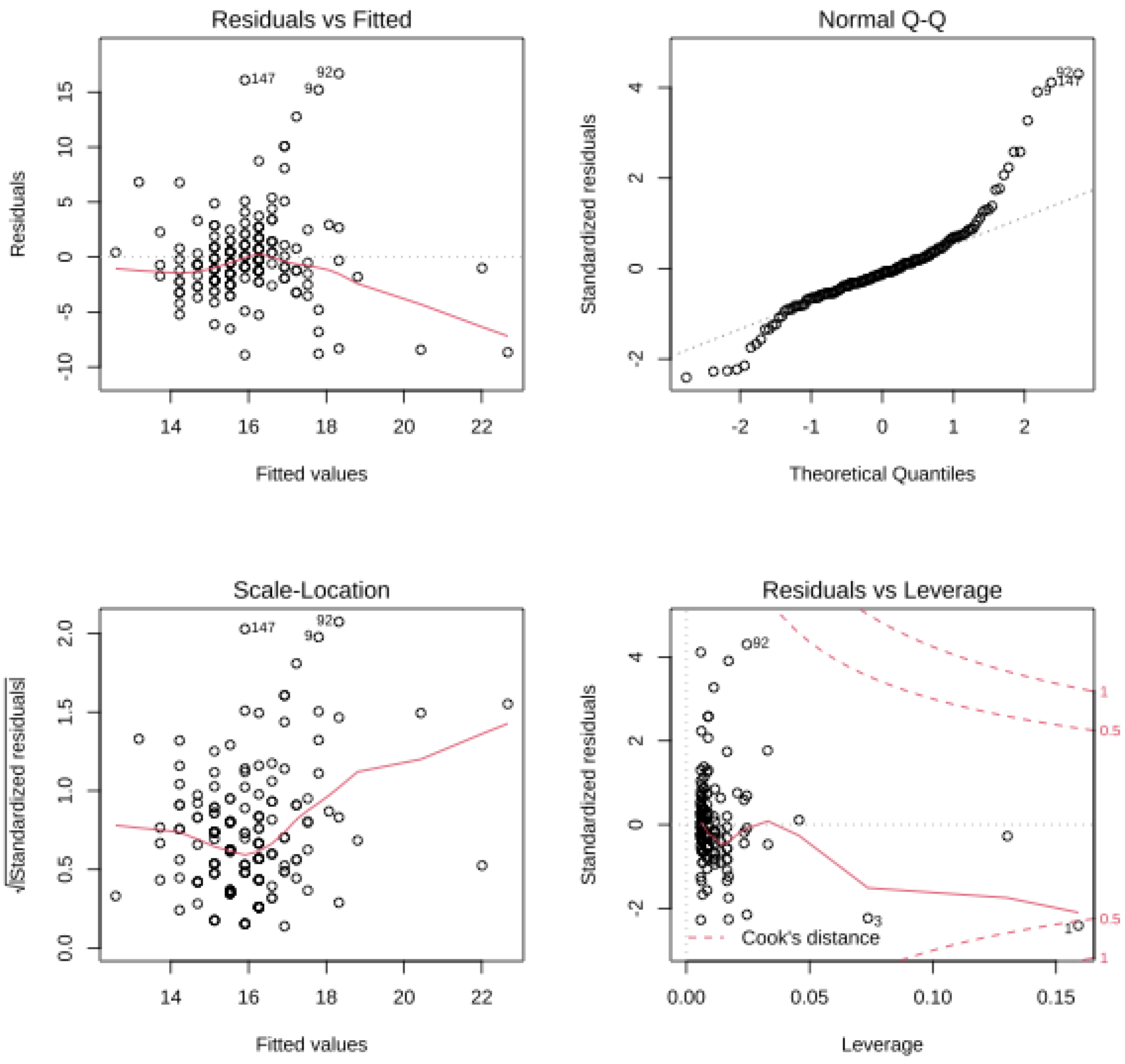
Diagnostic plots

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3002 -1.6629  0.0412  1.8944  3.9775
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -38.45509    8.04901  -4.778  0.00139 **
## x             0.67461    0.05191  12.997 1.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.253 on 8 degrees of freedom
## Multiple R-squared:  0.9548,    Adjusted R-squared:  0.9491
## F-statistic: 168.9 on 1 and 8 DF,  p-value: 1.164e-06
```

```
setwd("/Users/mariacuellar/Documents/Penn/Classes/Crim 250 - Statistics for the Social Sciences/Assignments/

dat <- read.csv(file = 'dat.nsduh.small.1.csv')

plot(dat$mjage, log(dat$cigage))
```


Residuals vs Fitted
Residuals
Fitted values
-10
-5
0
5
10
15
14
16
18
20
22
147
92
9
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-2
0
2
4
-2
-1
0
1
2
92
147
9
Scale-Location
√|Standardized residuals|
Fitted values
0.0
0.5
1.0
1.5
2.0
14
16
18
20
22
147
92
9
Residuals vs Leverage
Standardized residuals
Leverage
-2
0
2
4
0.00
0.05
0.10
0.15
92
3
1
Cook's distance
1
0.5
0.5
1