Lecture 11: Linear regression in R

Criminology 250

Prof Maria Cuellar

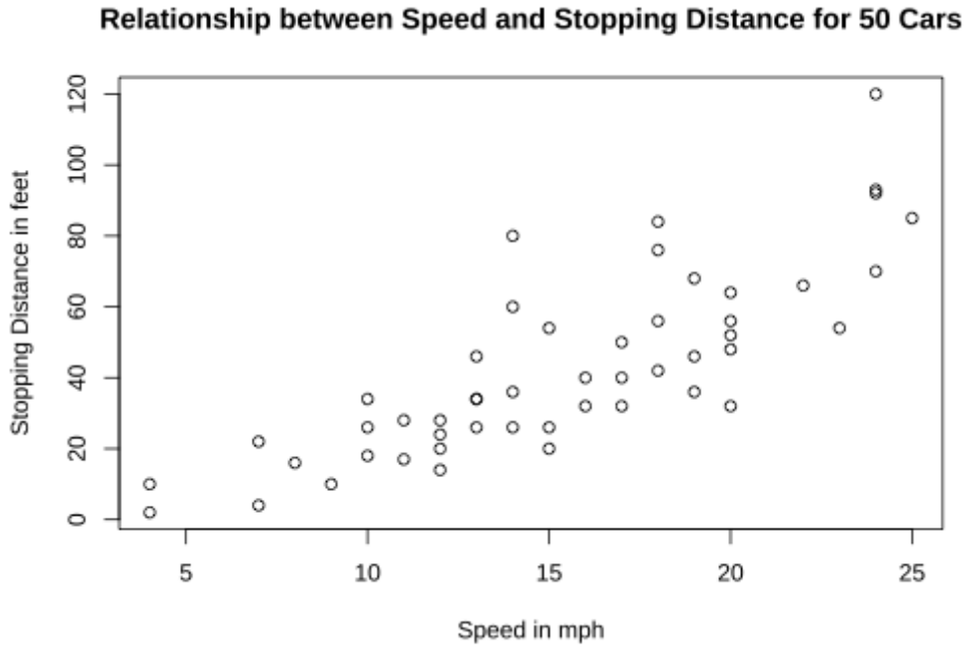
University of Pennslyvania

Linear regression

We will be using an example from the cars dataset in R.

This is what the scatterplot looks like

```
## [1] "speed" "dist"
```



Linear model

The model

- The linear regression model assumes that the *means* of the distributions of y's for each x fall along the line, even though the individuals are scattered around it.
- The model is

$$\mu_y = \beta_0 + \beta_1 x.$$

  We use Greek letters to denote idealized models. Whenever we use linear regression, we're assuming that this is actually how the data points are distributed.

- Are the data really going to be distributed like this?

Linear model

The errors

- No, not all the individual y's are at these means. Some are above, some below. So, like all models, this one makes errors. They are model errors so we call them ϵ.
- When we include the errors, we can actually say that each individual y is along the line, with some variation,

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where x and y are each individual point's x and y values, the βs are the parameters of the model (slope and intercept), and ϵ are the model errors that soak up the deviation from the model to the actual point. This equation is true for each data point.

Linear model

The regression line

- We estimate the βs by finding a regression line

$$\hat{y} = b_0 + b_1 x,$$

  as in the previous class. The residuals, $e = y - \hat{y}$ are the sample-based versions of the errors ϵ.

- We use a method called least squares regression, which minimizes the vertical distance from the data points to the regression line (the sum of the squares of the residuals), to get reasonable estimates of the parameters of this model from a random sample of data.

Assumptions

Linearity assumption

- This is satisfied if a scatterplot of x and y looks straight. If the true relationship between x and y is far from linear and we use a straight line to fit the data, our entire analysis will be useless, so we always check this first.
- How to check? You can see violations of this if you plot a scatterplot of the residuals against x or against the predicted values $\hat{y}$. That plot will have a horizontal direction and should have no pattern if the condition is satisfied.

Assumptions

Independence assumption

- The errors in the true underlying regression model (the ϵs) must be independent of each other. There is no way to check this is true. We check displays of the regression residuals for evidence of patterns, trends, or clumping, any of which would suggest a failure of independence. e.g. For time series, the error our model makes today may be similar to the one it made for yesterday.
- How to check? You can see violations of this by plotting the residuals against x and looking for patterns. Or, plot the residuals vs. the residuals offset or lagged by one time position. Neither plot should show patterns.

Assumptions

Equal variance assumption (homoscedasticity)

- The variability in y should be about the same for all values of x. The standard deviation of the residuals "pools" information across all of the individual distributions at each x-value, and pooled estimates are appropriate only when they combine information for groups with the same variance.
- How to check? A scatterplot of y against x offers a visual check. Be alert for a "fan" shape or other tendency for the variation to grow or shrink in one part of the scatterplot. Often, it is better to look at the residuals plotted against the predicted values $\hat{y}$.

Assumptions

Normal population assumption

- We assume the errors around the idealized regression line at each value of x follow a Normal model. Why?
  We need this assumption so we can use Student's t-model for inference.
- This assumption becomes less important as the same size grows because the model is about means and the Central Limit Theorem takes over.
- How to check? Nearly normal condition (qq plot) and outlier condition (Cook's distance).

Important note

We don't expect the assumptions to be exactly true, and we know that all models are wrong, but the linear model is often close enough to be very useful.

Procedure for linear regression

In regression, there's a little catch. The best way to check many of the conditions is with the residuals, but we get the residuals only *after* we compute the regression. Before we compute the regression, however, we should check at least one of the conditions.

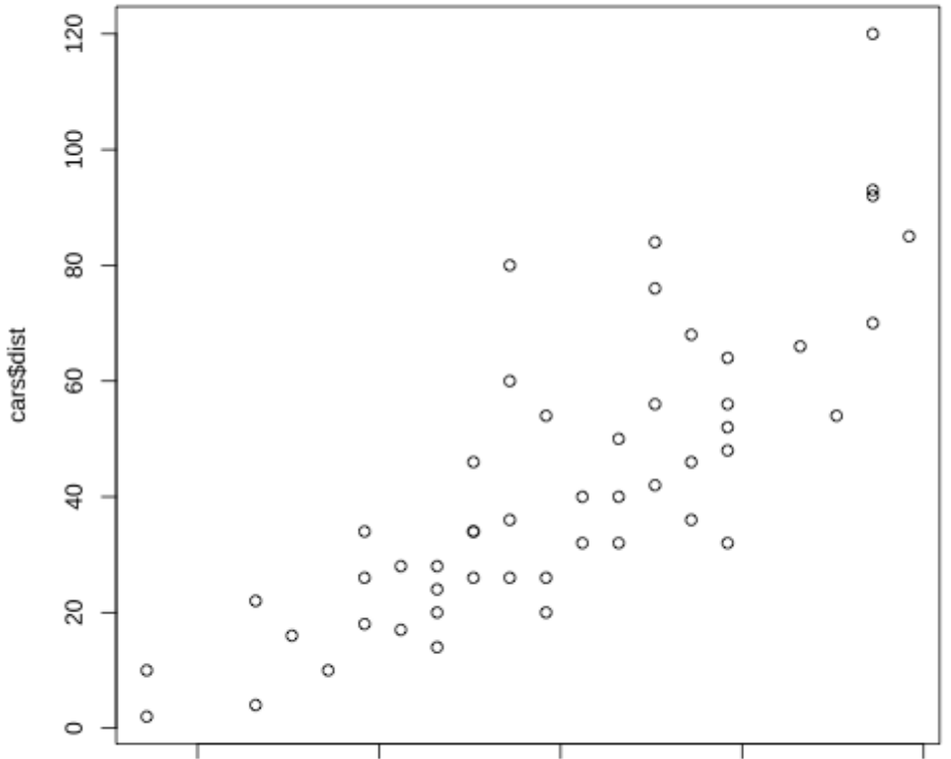
So we work in this order:

1. Make a scatterplot of the data to check the Straight Enough Condition. (If the relationship is curved, try re-expressing the data. Or stop.)
2. If the data are straight enough, fit a regression and find the predicted values, $\hat{y}$, and the residuals, e.
3. Make a scatterplot of the residuals against x or against the predicted values. This plot should have no pattern. Check in particular for any bend (which would suggest that the data weren't all that straight after all), for any thickening (or thinning), and, of course, for any outliers. (If there are outliers, and you can correct them or justify removing them, do so and go back to step 1, or consider performing two regressions—one with and one without the outliers.)
4. If the data are measured over time, plot the residuals against time to check for evidence of patterns that might suggest they are not independent.
5. If the scatterplots look OK, then make a histogram and Normal probability plot of the residuals to check the Nearly Normal Condition.
6. If all the conditions seem to be reasonably satisfied, go ahead with inference.

R linear regression formula and output

```
##
## Call:
## lm(formula = dist ~ speed.c, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.9800     2.1750  19.761  < 2e-16 ***
## speed.c       3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```
##   speed dist speed.c
## 1     4    2   -11.4
## 2     4   10   -11.4
## 3     7    4    -8.4
## 4     7   22    -8.4
## 5     8   16    -7.4
## 6     9   10    -6.4
```



`lm` R linear model output

The model above is achieved by using the `lm()` function in R and the output is called using the `summary()` function on the model.

- Formula call: The first item shown in the output is the formula R used to fit the data. Note the simplicity in the syntax: the formula just needs the predictor (x) and the target/response variable (y), together with the data being used (dat).
- Residuals: The next item in the model output talks about the residuals. Residuals are essentially the difference between the actual observed response values (distance to stop dist in our case) and the response values that the model predicted.
- Coefficient - Estimate: The coefficient Estimate contains two rows; the first one is the intercept. The intercept, in our example, is essentially the expected value of the distance required for a car to stop when we consider the average speed of all cars in the dataset. In other words, it takes an average car in our dataset 42.98 feet to come to a stop. The second row in the Coefficients is the slope, or in our example, the effect speed has in distance required for a car to stop. The slope term in our model is saying that for every 1 mph increase in the speed of a car, the required distance to stop goes up by 3.9324088 feet.

lm R linear model output

- Coefficient - Standard Error: The coefficient Standard Error measures the average amount that the coefficient estimates vary from the actual average value of our response variable. We'd ideally want a lower number relative to its coefficients. In our example, we've previously determined that for every 1 mph increase in the speed of a car, the required distance to stop goes up by 3.9324088 feet. The Standard Error can be used to compute an estimate of the expected difference in case we ran the model again and again. In other words, we can say that the required distance for a car to stop can vary by 0.4155128 feet. The Standard Errors can also be used to compute confidence intervals and to statistically test the hypothesis of the existence of a relationship between speed and distance required to stop.

- Coefficient - t-value: The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null hypothesis - that is, we could declare a relationship between speed and distance exist. In our example, the t-statistic values are relatively far away from zero and are large relative to the standard error, which could indicate a relationship exists. In general, t-values are also used to compute p-values.

- Coefficient - Pr(>t): The Pr(>t) acronym found in the model output relates to the probability of observing any value equal or larger than t. A small p-value indicates that it is unlikely we will observe a relationship between the predictor (speed) and response (dist) variables due to chance. Typically, a p-value of 5% or less is a good cut-off point. In our model example, the p-values are very close to zero. Note the 'signif. Codes' associated to each estimate. Three stars (or asterisks) represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude

lm R linear model output

- Residual Standard Error: Residual Standard Error is measure of the quality of a linear regression fit. Theoretically, every linear model is assumed to contain an error term E. Due to the presence of this error term, we are not capable of perfectly predicting our response variable (dist) from the predictor (speed) one. The Residual Standard Error is the average amount that the response (dist) will deviate from the true regression line. In our example, the actual distance required to stop can deviate from the true regression line by approximately 15.3795867 feet, on average. In other words, given that the mean distance for all cars to stop is 42.98 and that the Residual Standard Error is 15.3795867, we can say that the percentage error is (any prediction would still be off by) 35.78%. It's also worth noting that the Residual Standard Error was calculated with 48 degrees of freedom. Simplistically, degrees of freedom are the number of data points that went into the estimation of the parameters used after taking into account these parameters (restriction). In our case, we had 50 data points and two parameters (intercept and slope).

- Multiple R-squared, Adjusted R-squared: The R-squared (R2) statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. R2 is a measure of the linear relationship between our predictor variable (speed) and our response / target variable (dist). It always lies between 0 and 1 (i.e.: a number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable). In our example, the R2 we get is 0.6510794. Or roughly 65% of the variance found in the response variable (dist) can be explained by the predictor variable (speed). Step back and think: If you were able to choose any metric to predict distance required for a car to stop, would speed be one and would it be an important one that could help explain how distance would vary based on speed? I guess it's easy to see that the answer

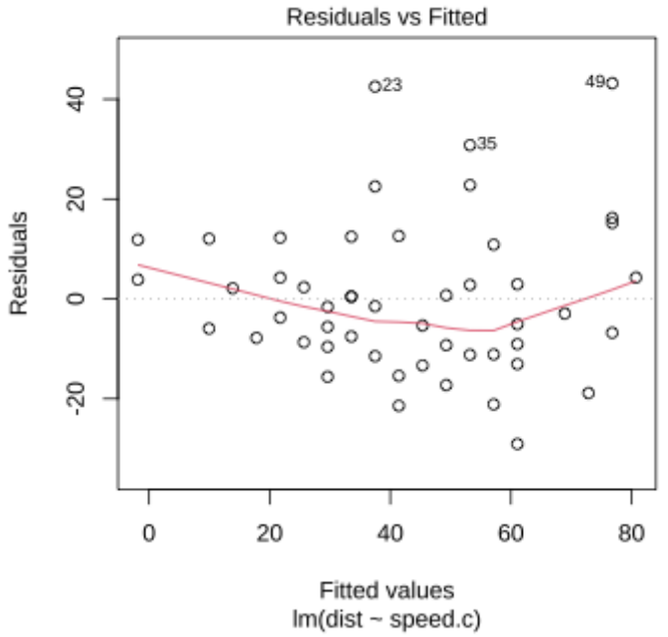
lm R linear model output

- F-Statistic: F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables. The further the F-statistic is from 1 the better it is. However, how much larger the F-statistic needs to be depends on both the number of data points and the number of predictors. Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis (H0 : There is no relationship between speed and distance). The reverse is true as if the number of data points is small, a large F-statistic is required to be able to ascertain that there may be a relationship between predictor and response variables. In our example the F-statistic is 89.5671065 which is relatively larger than 1 given the size of our data.

Diagnostics revisited

Residuals vs. fitted

```
plot(reg.output, which=1)
```



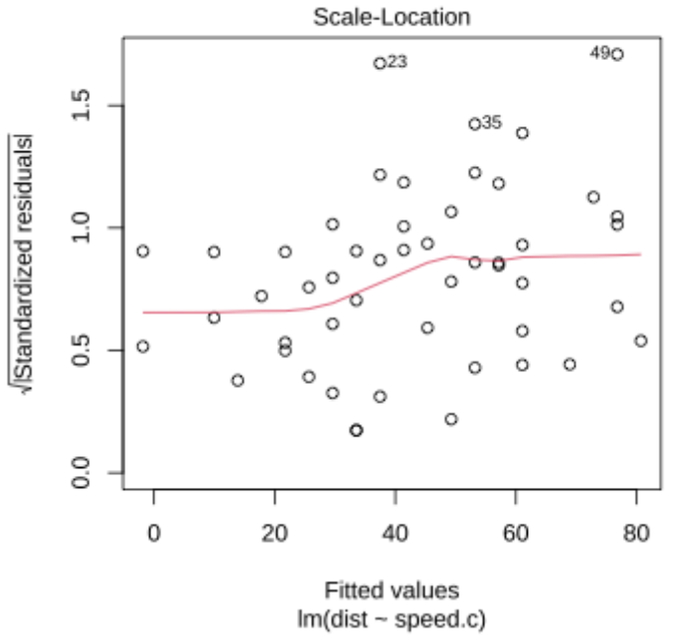
Residuals vs. fitted

You can think of the residuals as being estimates of the error terms. So anytime we're looking at a plot that involves residuals, we're doing so because we're trying to assess whether some assumption about the errors appears to hold in our data.

Looking at the Residuals vs Fitted plot (showing residuals on the y-axis and fitted y's on the x-axis), we see that the red line (which is just a scatterplot smoother, showing the average value of the residuals at each value of fitted value) is quite flat. This tells us that there is no discernible non-linear trend to the residuals. Furthermore, the residuals appear to be equally variable across the entire range of fitted values. There is no indication of non-constant variance.

Scale-location plot

```
plot(reg.output, which=3)
```

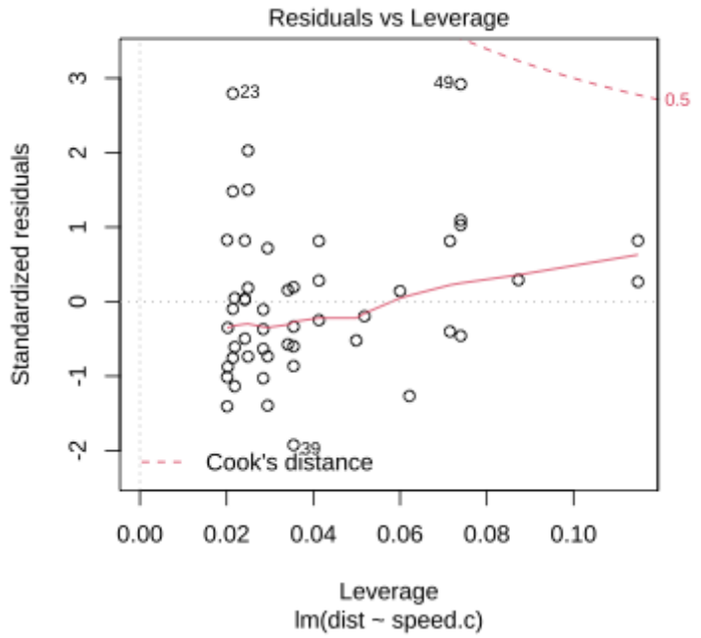


Scale-location plot

The scale-location plot is a more sensitive approach to looking for deviations from the constant variance assumption. If you see significant trends in the red line on this plot, it tells you that the residuals (and hence errors) have non-constant variance. That is, the assumption that all the ϵi have the same variance is not true. When you see a flat line like what's shown above, it means your errors have constant variance, like we want to see.

Outliers and the residuals vs. Leverage plot

```
plot(reg.output, which=5)
```

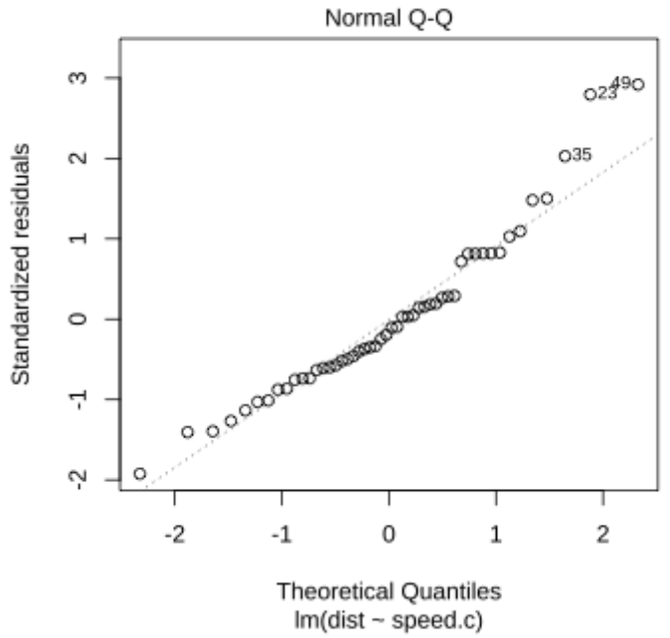


Outliers and the residuals vs. Leverage plot

There’s no single accepted definition for what consitutes an outlier. One possible definition is that an outlier is any point that isn’t approximated well by the model (has a large residual) and which significantly influences model fit (has large leverage). This is where the Residuals vs Leverage plot comes in.

Normal QQ plot

```
plot(reg.output, which=2)
```



Normal QQ plot

The Normal QQ plot helps us to assess whether the residuals are roughly normally distributed. If the residuals look far from normal we may be in trouble. In particular, if the residual tend to be larger in magnitude than what we would expect from the normal distribution, then our p-values and confidence intervals may be too optimisitic. i.e., we may fail to adequately account for the full variability of the data.

For next time:

- Central Limit Theorem
- Student's t-test

Interpretation

- R^2:
- Standard error for the slope is affected by: the spread around the line, the spread of x values, and the sample size:

$$SE(b_1) = \frac{s_e}{\sqrt{n-1}s_x}.$$

- Inference for the slope:

$$ \frac{b*1*-*beta_1}{SE(b_1)} *sim t*{n-2}, $$ where \(b_1\) is the estimated slope and \(\beta_1\) is the slope of the idealized regression line.

- Similarly for the intercept:

$$ \frac{b*0* - *beta_0}{SE(b_0)} *sim t*{n-2}. $$

- DVB Chp 25
- https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R
- https://www.andrew.cmu.edu/user/achoulde/94842/homework/regression_diagnostics.html