Lecture 3 - Review of Multivariate Regression

Samuel DeWitt, Ph.D.

Review of Multivariate Regression

What is regression?

Do you remember slopes from calculus in high school?

Perhaps the phrase ‘rise over reach’?

Review of Multivariate Regression

We use multivariate regression to understand the relationships between two or more variables.

The logic behind this is that, if we ‘control’ for the influence of other variables, we can isolate the *unique* influence of another variable.

However, there are some important caveats to this that we will talk about in this lecture.

Review of Multivariate Regression

Today, we will review correlation coefficients and *bivariate* regression, then move onto considering more than two variables at once.

First, I want to introduce scatterplots - the first stage in assessing a bivariate relationship between two continuous variables.

Scatterplots

A scatterplot is a data visualization for depicting the co-variation across two continuous, ratio-level variables.

By way of review, a continuous, ratio-level variable is one that:

1) has a meaningful zero point (e.g., 0 for an number of arrests variable indicates the person has not been arrested)
2) does not have a strict upper (or lower) bound (i.e., it can continue toward ∞)
3) movement from one value to another is in fixed units (e.g., a change from 1 to 2 is the same increase as a change from 99 to 100)

Scatterplots

A quick note: Interval level variables may also be included in scatterplots but they do not have a meaningful zero point (and examples of interval-level variables do not abound).

Scatterplots

A scatterplot boils down to the joint plotting X and Y coordinates for two variables.

In order for a scatterplot to make sense, movement along either axis must be meaningful (i.e., no nominal variables allowed!)

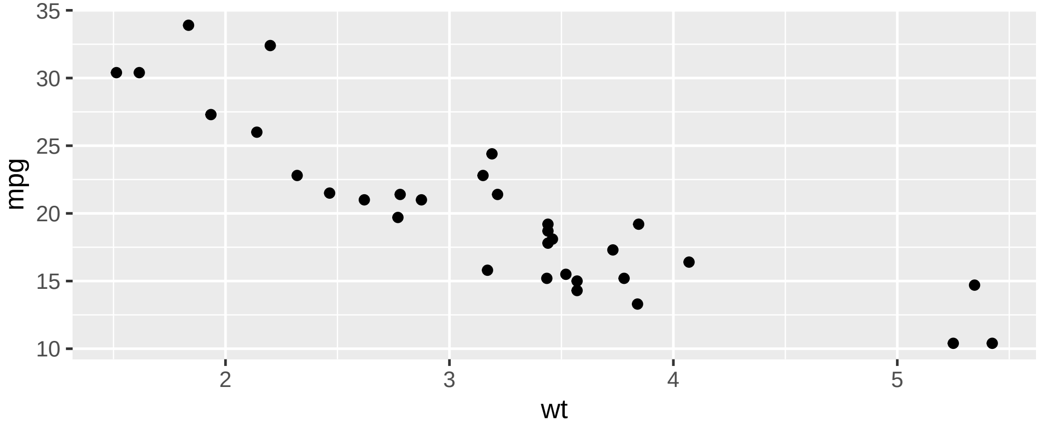
Scatterplots

Here's an example of a scatterplot using the mtcars data set that comes pre-packaged with R.

We want to know if the weight of a vehicle (wt) is correlated with its fuel efficiency (mpg).

Scatterplots

```
ggplot(mtcars,aes(x=wt,y=mpg))+geom_point()
```



Scatterplots

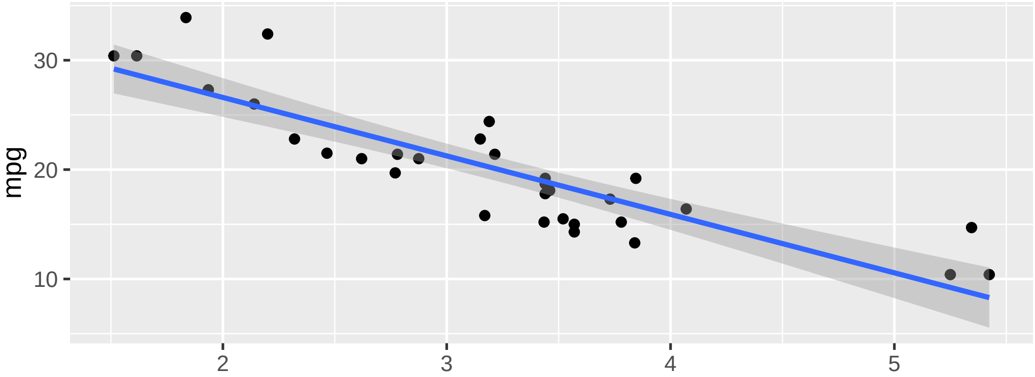
What does the relationship between a vehicle's weight and miles per gallon appear to be, just judging by this scatterplot?

There are ways to incorporate an estimate of this relationship into the plot using a linear trend (just like a bivariate regression would).

Scatterplots

```r
ggplot(mtcars,aes(x=wt,y=mpg))+geom_point() +
  geom_smooth(method="lm", se=TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Scatterplots

The line is fit using the geom_smooth() option within the ggplot() function.

I specify that the line is fit from a linear model ("lm") which means that the graph depicts the linear relationship between vehicle weight and fuel efficiency.

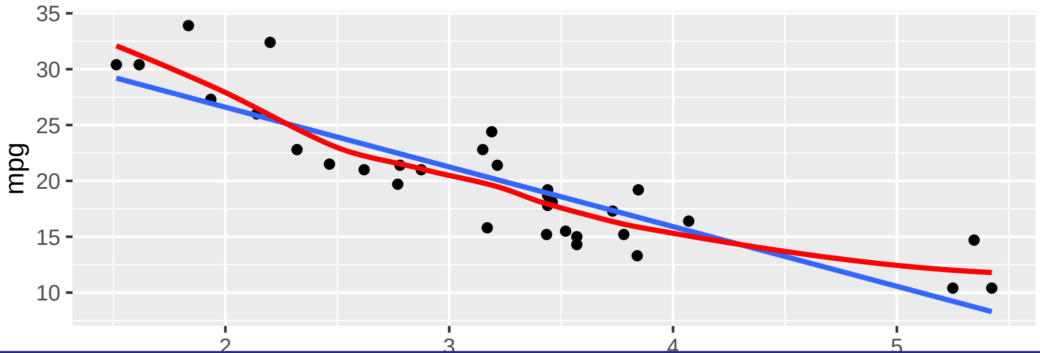
There are other lines we could fit, including a LOESS (locally estimated scatterplot smoothing).

A LOESS smoother can be preferred when data are sparse in certain regions and you worry about extrapolating too much from limited data points.

Scatterplots

```
ggplot(mtcars,aes(x=wt,y=mpg))+geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  geom_smooth(method="loess", se=FALSE, colour="red")
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Scatterplots

The blue line is the linear trend, while the red line is the LOESS smoother.

They are pretty similar, but differ the most in two areas. Where do they differ?

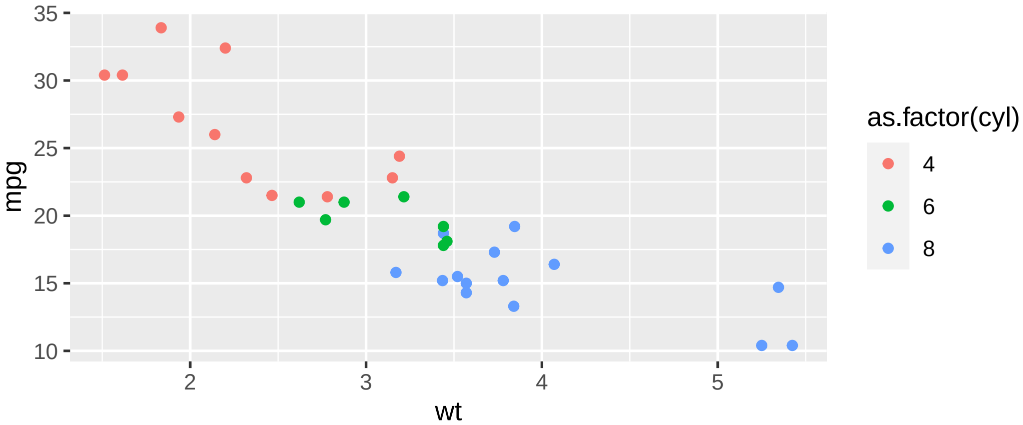
Scatterplots

What if we want to account for a third variable?

Well, there are ways to do this, but it will not truly account for the influence of this variable in the apparent impact of vehicle weight on fuel efficiency.

Scatterplots

```
ggplot(mtcars,aes(x=wt,y=mpg,colour=as.factor(cyl)))+geom_point()
```



Scatterplots

In this example, I changed the color of the plot points using the colour() option.

The relationship between vehicle weight and fuel efficiency appears to be slightly different depending upon the number of cylinders a car has. How is it different?

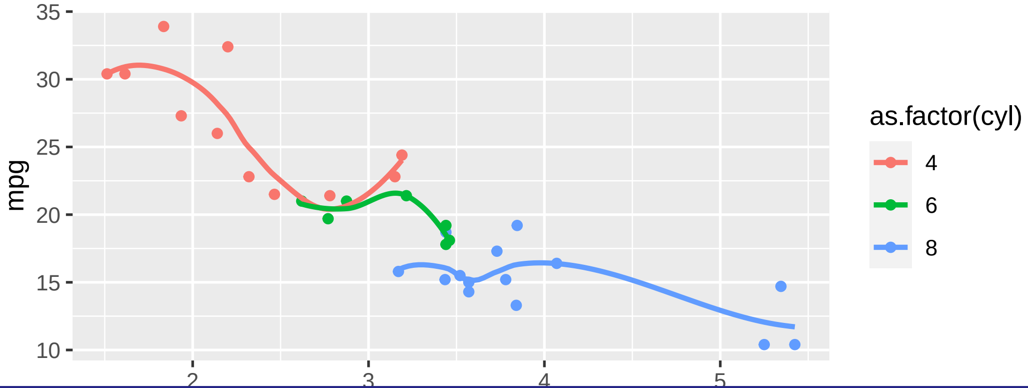
Scatterplots

As a final example, I will show you that you can fit conditional LOESS lines based upon values of a third variable.

Scatterplots

```
ggplot(mtcars,aes(x=wt,y=mpg,colour=as.factor(cyl)))+geom_point() +
  geom_smooth(method="loess", se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Scatterplots

The relationship certainly appears like it may be more complicated than the original, simple downward linear trend.

Ideally, we want to know the unique influence of a vehicle's weight **controlling** for its number of cylinders. This would require estimating a multivariate regression equation.

Before we do that, we need to build the intuition for that model a little more through a review of correlation coefficients.

Correlation Coefficients

A correlation “standardizes” the association between two variables.

Correlations range from -1 to 1, with values closer to either boundary indicating a stronger association.

Correlations of absolute value from:

1) .00 to .30 indicate a *weak* relationship
2) .31 to .50 indicate a *moderate* relationship
3) .51 to 1.00 indicate a *strong* relationship

They also tell us the **direction** of the relationship

Correlation Coefficients

A correlation coefficient is represented by the r symbol, and is calculated as follows:

$$\begin{aligned} r &= \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum (X-\overline{X})^2 \sum (Y-\overline{Y})^2}} \\ &= \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X^2)][n\sum Y^2 - (\sum Y^2)]}} \\ &= \frac{\sum XY - n\overline{XY}}{\sqrt{[\sum X^2 - n\overline{X}^2][\sum Y^2 - n\overline{Y}^2]}} \end{aligned} \tag{1}$$

Correlation Coefficients

We need to calculate the variance of X, the variance of Y, and the crossproduct (or *covariance*) of X and Y.

But, I'll let you guys use the built-in R commands to do this.

Here are some examples using the mtcars data.

Correlation Coefficients

```r
cor(mtcars$wt,mtcars$mpg)
```

```
## [1] -0.8676594
```

```r
cor(mtcars$cyl,mtcars$mpg)
```

```
## [1] -0.852162
```

```r
cor(mtcars$cyl,mtcars$wt)
```

```
## [1] 0.7824958
```

What should I infer from these values about the bivariate relationships between these pairs of variables?

Correlation Coefficients

An additional benefit of correlation coefficients is that they can be directly compared, since they reflect standardized-unit increases in one variable as it relates to a one standardized unit increase in the other variable.

Specifically, we can interpret the coefficient as such: a one standard-deviation unit increase in X is **associated** with a "###" standard-deviation unit increase in Y. Fill in the blanks here assuming weight is X and fuel efficiency is Y.

Given this, which of those relationships would we call the *strongest*?

Correlation Matrices & Correlograms

But what if we want to display all the relevant correlations between the different variables in a data set?

We could create a correlation matrix where the rows and columns each represent a single variable in the data.

Here's what one such correlation matrix would look like. For this example you will need to install the "corrplot" package.

You will also need to run the following line of code:
source("http://www.sthda.com/upload/rquery_cormat.r")

Correlation Matrices & Correlograms

```r
rquery.cormat(mtcars, graph=FALSE)
```

```
## $r
##        carb    wt    hp   cyl  disp  qsec   vs  mpg drat   am gear
## carb      1
## wt     0.43     1
## hp     0.75  0.66     1
## cyl    0.53  0.78  0.83     1
## disp   0.39  0.89  0.79   0.9     1
## qsec  -0.66 -0.17 -0.71 -0.59 -0.43     1
## vs    -0.57 -0.55 -0.72 -0.81 -0.71  0.74    1
## mpg   -0.55 -0.87 -0.78 -0.85 -0.85  0.42 0.66    1
## drat -0.091 -0.71 -0.45  -0.7 -0.71 0.091 0.44 0.68    1
## am    0.058 -0.69 -0.24 -0.52 -0.59 -0.23 0.17  0.6 0.71    1
```

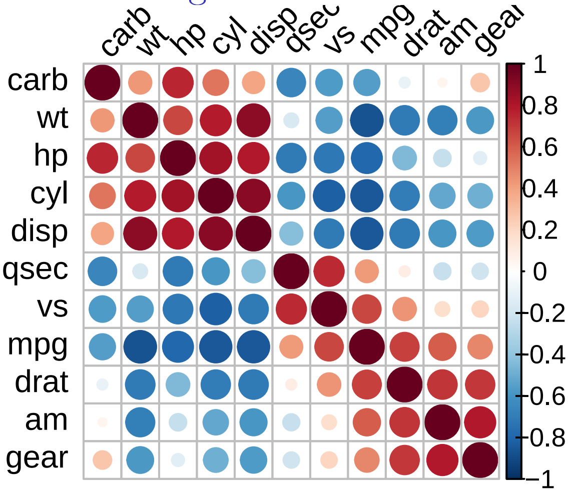
Correlation Matrices & Correlograms

That's messy, to say the least.

Notice, though, the diagonal line of 0s - these are omitted correlations because a variable cannot be correlated with itself.

Let's strictly look at a plot and see if that helps us make more sense of the distributions between variables in these data.

Correlation Matrices & Correlograms



Correlation Matrices & Correlograms

That's a bit easier to make sense of - note that the strength of the correlation is represented by the the color density and direction of the relationship by blue (negative) and red (positive).

Note that I also specified "full" for the correlation matrix display - this simply displays both the upper and lower parts of the correlation matrix (which are identical).

What strong positive correlations do you notice? Negative correlations?

Bivariate Regression

Bivariate regression takes the correlation coefficient one step further.

It can tell us the change we expect to observe in the dependent variable for every **one-unit change** in the independent variable.

Bivariate Regression

Population regression equations take the following general form:

$$Y = \alpha + \beta X + \epsilon$$

Where α and β are population **parameters** that summarize the relationship between X and Y.

ϵ represents the errors you make using this equation to predict values of Y

Bivariate Regression

This same equation, but for sample (which we almost always have) is the following:

$$Y = a + bX + e$$

There are some slight differences:

1) a represents the Y-intercept, and is a sample estimate for α
2) b represents the slope, and is the sample estimate for β
3) e represents the errors of the regression equation, and is the sample estimate for ϵ

Note: Y may also be represented with $\hat{Y}$ - this simply means it is a *predicted* value.

Bivariate Regression - A Quick Aside on Population v. Samples

Why the distinction between population and sample equations?

Well, even though the terms in the equation are the same, we use different symbols to reflect the fact that, since this is a sample, we should assume there is **sampling error** in our estimates.

Bivariate Regression - A Quick Aside on Population v. Samples

Sampling error reflects the fact that we could have taken an infinite amount of different samples.

Therefore, we would fully expect our estimates to have some variability over these samples.

By contrast, α, β, and ϵ are not assumed to vary - they are the true population **parameters**.

Calculating a Bivariate Regression

The equation for a bivariate regression is quite similar to that for a correlation coefficient:

$$b = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2} = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2} = \frac{\sum XY - n\overline{XY}}{\sum X^2 - n\overline{X}^2}$$

Calculating a Bivariate Regression

Luckily for you all, I will not have you calculate this by hand (you should have done that in your Stats class, anyhow).

Instead, I will have you use the lm() function in R - this stands for **linear model**.

This function has a very particular form to it, so I will demonstrate its application using the mtcars data below.

Estimating a Bivariate Regression in R

The function in R takes the following general form: *lm(formula, data)*

The *formula* part refers to the which variables R should treat as the outcome and predictor.

In a bivariate regression, there are only one of each.

In a multivariate regression, there can be two or more predictor variables, but still only one outcome.

Estimating a Bivariate Regression in R

The *formula* is written generally as: outcome ∼ predictor.

To incorporate additional predictors you can just add them after a + sign:

$$\text{outcome} \sim \text{predictor1} + \text{predictor2} + \ldots$$

Estimating a Bivariate Regression in R

The *data* part merely tells R which data frame it needs to search in for the predictor and outcome variable data.

This section must be specified if you do not want to write mtcars$ in front of every single variable name you want to use.

I assure you that you don't want to do that.

Note - sometimes R makes you do it anyway, though (you'll see an example later).

Estimating a Bivariate Regression in R

A pretty simple example:

```r
lm(mpg~wt, data=mtcars)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Coefficients:
## (Intercept)           wt
##      37.285       -5.344
```

Who wants to attempt to provide an interpretation for each of those coefficients? Volunteers first, then I'll just cold call on you!

Estimating a Bivariate Regression in R

The intercept, or a from the sample equation, is 37.285. This is the value of Y when the regression line crosses the Y-axis.

Since this is generally at a value of 0 for X, this means that 37.285 miles per gallon is what we would expect from a car that weights zero pounds.

That's. . . .not very helpful.

Estimating a Bivariate Regression in R

The intercept can often represent some nonsensical value of Y that we would never expect to observe in practice.

This is because the intercept is only an *anchoring point* of the regression line. It can change quite a bit when additional variables are added to the equation.

One way to get a meaningful intercept is to **de-mean** the predictor variables. This is accomplished by subtracting the mean of the predictor from each observation.

Estimating a Bivariate Regression in R

What about the coefficient for weight? What does -5.344 mean?

Well, this means that, for every **one unit** increase in the weight variable we would expect, *on average* a 5.344 unit decrease in miles per gallon.

That seems like a really large coefficient! Its size is a bit deceiving though - what does a *one unit* change in weight represent, do you think?

Estimating a Bivariate Regression in R

That's not all that the lm() function can produce though.

If we create an **object** out of the regression model it will store additional pieces of information.

Storing a regression model as an object in R is as simple as storing any other object in R - you just need to give it a name and use the assignment operator (<-).

Estimating a Bivariate Regression in R

```r
model1<-lm(mpg~wt, data=mtcars)
```

Estimating a Bivariate Regression in R

Notice how it doesn't show any output? You can get the output to display by typing in the new object's name as such:

```
model1
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Coefficients:
## (Intercept)           wt
##      37.285       -5.344
```

Estimating a Bivariate Regression in R

What I also want you to notice is that an object called **model1** is now in your working environment!

There's also a litte blue arrow button next to it - if you click on that button it will display a list of all the additional pieces of information the linear model produces (it's a long list).

This list will come in handy later when we run some regression **diagnostics**, or checks to make sure our model is not violating any of the required assumptions of regression.

Estimating a Bivariate Regression in R

You can also get additional information about linear model using the summary() command:

```r
summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Interpreting a Bivariate Regression in R

That's a lot more information to interpret, but all of it is helpful.

Let's go over each item.

First, the **Residuals** - a *residual* is another term for error - it's the difference between what the regression predicts for the value of Y and what value we actually observe.

Interpreting a Bivariate Regression in R

Second, the **Coefficients** section - there are four columns:

1) Estimate
2) Std. Error (Standard Error)
3) t value
4) Pr(>|t|) (Probability)

Interpreting a Bivariate Regression in R

1) Estimate

The estimate is merely the coefficient or, in the case of the intercept, the value of Y when X is zero.

Interpreting a Bivariate Regression in R

2) Std. Error (Standard Error)

The standard error reflects the degree of variation we expect in the coefficient estimates.

In simple terms, it reflects the degree of precision which we feel the coefficient estimate (a sample **statistic**) is a reflection of the population **parameter*.

Interpreting a Bivariate Regression in R

3) t value

This is the ratio of the coefficient to its standard error. It allows us to plot the coefficient along a probability distribution.

Plotting the coefficient estimate along a probability distribution allows us to determine the likelihood of achieving that result.

Interpreting a Bivariate Regression in R

4) Pr(>|t|) (Probability)

This is obtained from a probability distribution after you have obtained some standardized statistic to represent the coefficient. In this case, the t value is placed along a probability distribution (Student's t distribution) to be exact.

Placement along the distribution is indicative of how likely or unlikely it would be to obtain a coefficient estimate like the one we have if the **null hypothesis** is true.

In general, the **null hypothesis** is that there the **sample statistic** has a score of 0 (or near that) along the probability distribution.

Interpreting a Bivariate Regression in R

Residual standard error is the square root of the residual **sum of squares** divided by the residual degrees of freedom.

This tells us the error we expect to make in our predictions for a *typical* case in the data.

Interpreting a Bivariate Regression in R

Multiple R-Squared and **Adjusted R-Squared** reflect the amount of variation in the outcome that is *explained* by the predictor variables.

The difference between the two is that the **Adjusted R-Squared** is corrected for degrees of freedom, and will always be a more conservative estimate of the true population value.

A value of 0.7528 can be interpreted as a vehicle's weight explaining 75.28% of the variation in its miles per gallon.

A Note on R-Squared

A great deal is often made of this statistic, though the reasons why are questionable.

R-Squared cannot tell you whether the relationship is causal, or if it is spurious.

Therefore, it's a helpful *indicator* that your regression is explaining *enough* of the outcome, but not that your primary independent variable(s) are truly causing that variation.

Interpreting a Bivariate Regression in R

The **F-statistic** indicates that the included predictor variables contribute significantly toward explaining variation in the outcome variable.

Strictly speaking, it is the ratio of the **mean regression sum of squares** divided by the **mean error sum of squares**. You should recall the term **sum of squares** from your prior statistics classes.

The F-Statistic is a global test telling us whether all predictor variable coefficients are 0 (the **null hypothesis**) or not.

Plotting the Bivariate Regression

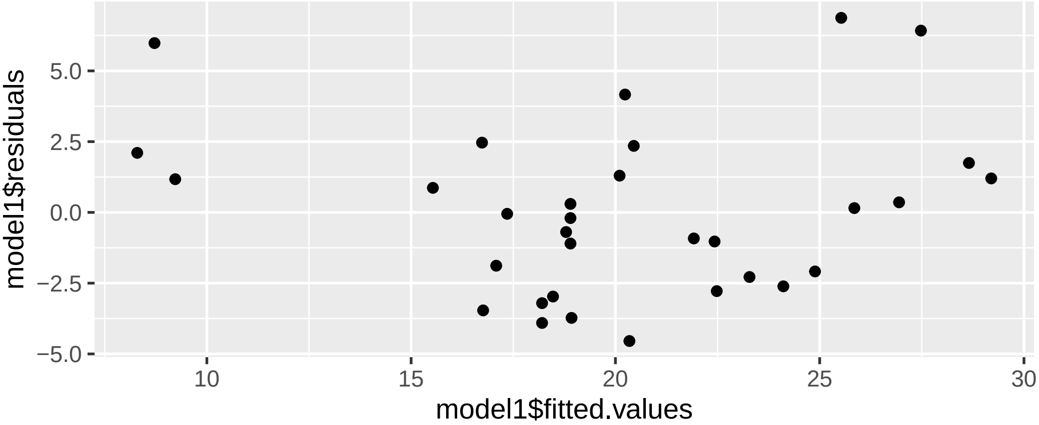
There are numerous elements of a bivariate regression we can visualize.

We have already done one - a scatterplot.

But we can also take a look at **residual** plots which tell us the extent to which our guesses are accurate and/or in what direction we typically make our errors.

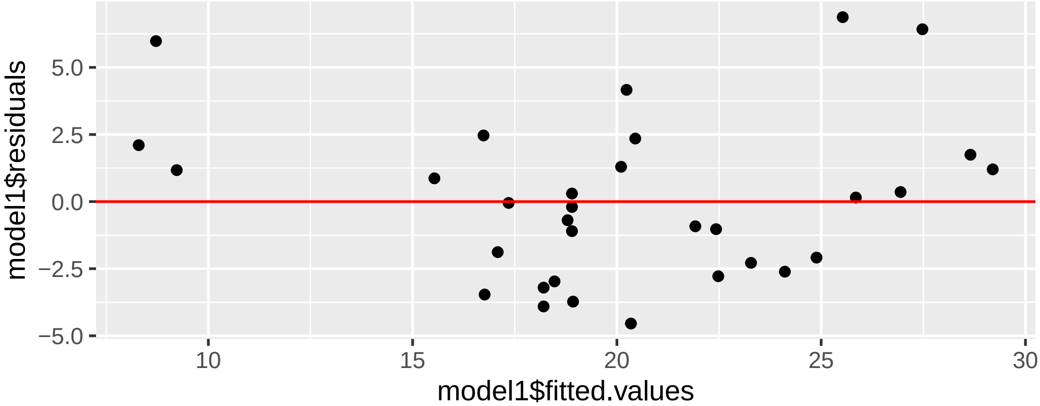
Plotting the Bivariate Regression

```
ggplot(model1, aes(x=model1$fitted.values,y=model1$residuals)) + geom_point()
```



Plotting the Bivariate Regression

```
ggplot(model1, aes(x=model1$fitted.values,y=model1$residuals)) + geom_point() +
  geom_hline(yintercept=0, colour="red")
```



Plotting the Bivariate Regression

What do you think we are looking for here?

Plotting the Bivariate Regression

What do you think we are looking for here?

That's right - we want there to be balance in the number and magnitude of the errors below and above the line at 0.

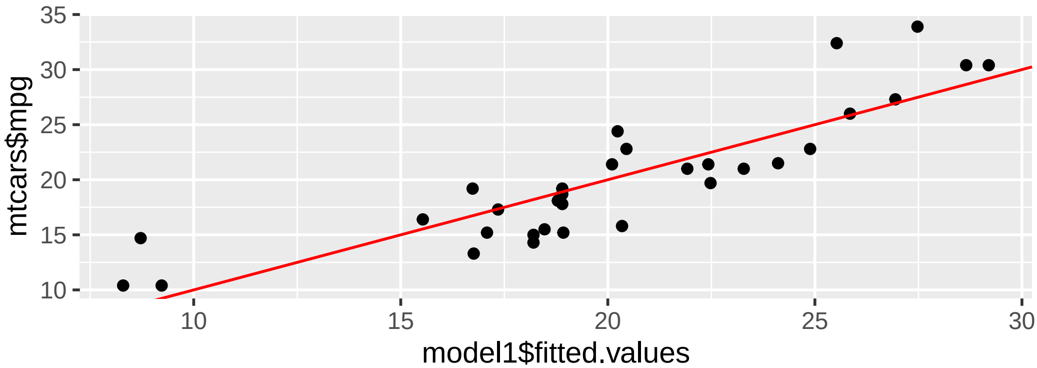
The line at 0 means we made no error in our prediction, while any deviation from that indicates the difference between the value we predicted for a case and its actual value.

Are the errors balanced here? What might be a sign of imbalance?

Plotting the Bivariate Regression

We can also simply plot predicted versus actual values to see these differences.

```r
ggplot(,aes(y=mtcars$mpg,x=model1$fitted.values)) + geom_point() +
  geom_abline(intercept=0, slope=1, colour="red")
```



Plotting the Bivariate Regression

The basic inference remains the same, but how it looks is a bit different. This is because we now have to infer the magnitude of errors from the disparity between the plots points.

Either orientation works, but I find they answer slightly different questions.

The first orientation indicates the overall normality of the residuals, while the second speaks to better to the **homoskedasticity** of the errors.

Estimating a Multivariate Regression

The last point has to do specifically with an assumption of OLS (**Ordinary Least Squares**) regression and is generally the point of one of several procedures called regression **diagnostics**.

These diagnostics are important in showing that *your* regression satisfies the assumptions of OLS, otherwise the model has one or more weaknesses that influence you ability to conduct inference.

We will talk more about these soon, but first we must review the procedure for adding a third (or fourth, or fifth. . .) variable to our OLS regression.

Estimating a Multivariate Regression

We have a new population and sample equation to estimate, though it's just slightly different from the bivariate case:

Population:

$$Y = \beta_0 X_1 + \beta_2 X_2 \ldots \beta_k X_k + \epsilon$$

Sample:

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots b_k X_k + e$$

Estimating a Multivariate Regression

Actually estimating this by hand is quite cumbersome, so we will use the lm() formula in R to do so. Basically, it amounts to calculating the correlations between each set of X variables as well as their respective correlations with Y.

If you recall the equation for the correlation coefficient from earlier - you would have to estimate this equation for each unique pair of variable in the data.

The formulas to compute the standard errors for the slopes (not shown above) become much more complicated, also.

Therefore, we are going to let R run the multivariate regressions for us (unless you anger me this semester, then it'll be on the first exam).

Estimating a Multivariate Regression

```r
model2<-lm(mpg~wt+cyl+am, data=mtcars)
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ wt + cyl + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## wt           -3.1251     0.9109  -3.431  0.00189 **
## cyl          -1.5102     0.4223  -3.576  0.00129 **
## am            0.1765     1.3045   0.135  0.89334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

Estimating a Multivariate Regression

As I stated earlier, it's quite easy to add more variables to the model.

This, however, complicates the interpretation of the coefficients in the model. What does the intercept mean now?

How would we interpret the coefficient for vehicle weight?

Are all the variables significantly related to fuel efficiency?

Assessing the Quality of a Multivariate Regression

Many researchers assess the quality of the model using the R-Squared value or F-Statistic.

I tend to think a regression is only as good as the assumptions it makes, and whether these assumptions are reasonable.

We will set aside causal assumptions for a moment and speak about the strictly technical assumptions of the regression model.

Assumptions of a Standard Linear Regression

There are six primary assumptions of a standard linear regression model (i.e., a multivariate ols).

1) The relationship between X and Y is linear (or linear in the parameters)
2) Non-stochasticity of X
3) Full rank
4) Errors have zero mean
5) Errors are i.i.d.
6) Errors are normally distributed

An assumption is something we hold to be true in order to **identify** the model. If the assumption is violated, this weakens the model, though some types of violations are less serious than others.

Assumptions of a Standard Linear Regression - 1) Linearity

We assume that the regression equation reflects a linear relationship between X and Y (at least roughly linear).

Or...that the relationship between X and Y can be expressed as a linear equation (linear in the parameters). What this means is that some function of Y or combination of functions of Y and X produce a linear relationship.

What kinds of functions do you think this means?

Assumptions of a Standard Linear Regression - 1) Linearity

Here are just some functions we might use to *express* a **non-linear** relationship between X and Y as *linear in the parameters*:

1) A polynomial model: $y = \alpha + \beta_1 X + \beta_2 X^2 + \epsilon$
2) An interaction model: $y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$
3) A log-linear model: $ln(y) = \alpha + \beta_1 X + \epsilon$
4) A semi-log model: $y = \alpha + \beta_1 ln(X) + \epsilon$
5) A log-log model: $ln(y) = \alpha + \beta_1 ln(X) + \epsilon$

Assumptions of a Standard Linear Regression - 1) Linearity

For models with polynomials, we are testing whether the relationship between X and Y differs at higher/lower levels of X.

A prime example is the age-crime curve. The relationship between age and the propensity to commit a crime is best expressed with 4 terms:

1) Age
2) Age^2
3) Age^3
4) Age^4

Note - when we have long-term data, that is.

Assumptions of a Standard Linear Regression - 1) Linearity

We could use four age variables (up to 4th power) because the age-crime curve has three **inflection points** where the slope between age and criminal propensity rapidly changes.

1) Early teens (rapid increase)
2) Late teens (rapid decline)
3) Early adulthood (line smooths out)

So, to satisfy the linearity assumption in the age-crim example, we need these polynomials.

Assumptions of a Standard Linear Regression - 1) Linearity

For models with interaction terms, we are asking whether the regression slope for one variable is moderated by another.

That is, does some other variable condition how steep the slope is for certain subsets of cases in our data?

Social support and stress is a good example - high stress and low social support can result in medical/psychological problems (high blood pressure, anxiety, etc...) while high stress and high social support tends not to. Why?

Assumptions of a Standard Linear Regression - 1) Linearity

For models with natural logs (the ln()) we are asking whether the distribution of the logged variables simply needs to be corrected for skew.

This is commonly used when there are a small number of *high* values in a predictor or outcome variable that could skew the analysis.

Good examples of typically skewed variables include: household income, number of arrests, and certain crime rates.

What do you think the new interpretation of the coefficient would be if we take the natural log of the variable?

Assumptions of a Standard Linear Regression - 2) Non-stochastic X

First, what does stochastic mean?

Assumptions of a Standard Linear Regression - 2) Non-stochastic X

First, what does stochastic mean?

Okay, so this means that variation in X is non-random. In the context of an experiment, the researcher adjusts values of X (treatment) themselves.

Without an experiment, we have to statistically **control** for other relevant variables that could affect Y (the outcome).

Assumptions of a Standard Linear Regression - 3) Full rank

1) X is of full column rank, rank(X) $= \boldsymbol{k+1}$ Where X is a matrix with a column of 1's for the constant and $\boldsymbol{k}$ additional columns for each regressor.

2) You must have as many observations $(\boldsymbol{n})$ as there are parameters $(\boldsymbol{k+1})$ in the model.

3) No regressor is a perfect linear combination of other regressors. I.e., no perfect correlations between predictors and, for dummy variables, one category has to be excluded.

Assumptions of a Standard Linear Regression - 4) Zero-mean Error

The *average* error has a value of zero.

OLS makes this true by definition.

Move along, nothing else to see here...

Assumptions of a Standard Linear Regression - 5) Errors i.i.d.

This means that errors from the model are *identically and indepently distributed* and is a two-part assumption.

Errors are **homoskedastic**: the distribution of errors is constant across all levels of predictor variables. I.e., we do not make better or worse predictions based upon the value of X

Errors are **independently distributed**: a prediction error for one unit is not related to making a prediction error for another unit (in presence or magnitude).

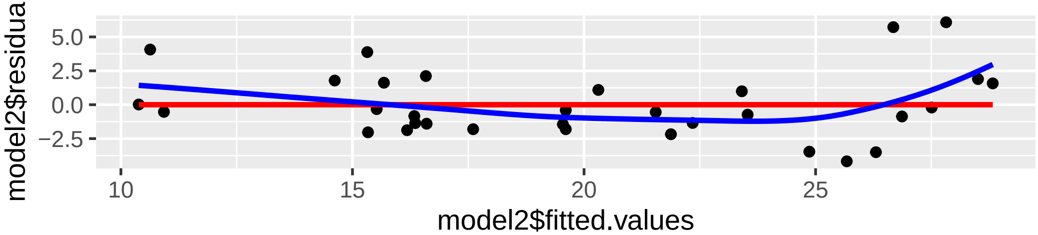
How can we test this assumption?

Assumptions of a Standard Linear Regression - 5) Errors i.i.d.

We can test this by plotting errors (residuals) against the predicted values of the dependent variable (LOESS and linear trends help with diagnosis!)

```
ggplot(,aes(x=model2$fitted.values, y=model2$residuals)) + geom_point()+
  geom_smooth(method='lm', color="red", se=FALSE)+
  geom_smooth(method='loess', color='blue', se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Assumptions of a Standard Linear Regression - 5) Errors i.i.d.

The other part of this assumption - independently distributed errors, is one we will tak about later this semester.

Basically, we need more complicated tests to determine if there's some type of autocorrelation in our errors, and we need some way to **cluster** observations to detect these patterns.

Assumptions of a Standard Linear Regression - 6) Normal Errors

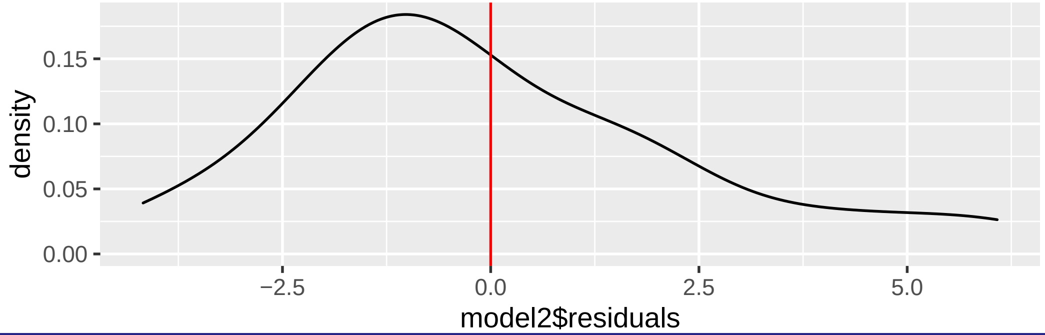
Distribution of errors assumes the (approximate) shape of the normal distribution (or we'd expect it to given a large enough sample size).

We may also appeal to the t-distribution here, but need to take into account our degrees of freedom for the regression $(n - k - 1)$.

Assumptions of a Standard Linear Regression - 6) Normal Errors

How do we assess normality? Density plots can help here. Are the errors normally distributed?

```
ggplot(,aes(x=model2$residuals)) + geom_density() +
  geom_vline(xintercept=0, color="red")
```


density
model2$residuals