Regression Lecture Notes[1]

Stanford GSE Math Camp 2015

Do Not Distribute Outside GSE

1 What is Regression Analysis?

- You're doing a research project (this will happen a lot) and you have a variable that you're particularly interested in understanding. This is your **dependent variable**.
- Additionally, you have information about a set of variables that you think are related to your dependent variable. These are your **independent variables** or **predictors**.
- Think of an example that's relevant for you.
  - What's your dependent variable?
  - What is a potential predictor?
  - What is your hypothesis for the relationship between these two variables?
- Regression gives us a framework to determine the nature of this relationship:
  - Test whether or not a **linear** relationship exists.
  - Quantify how strong that relationship is.

2 Visualizing Your Data

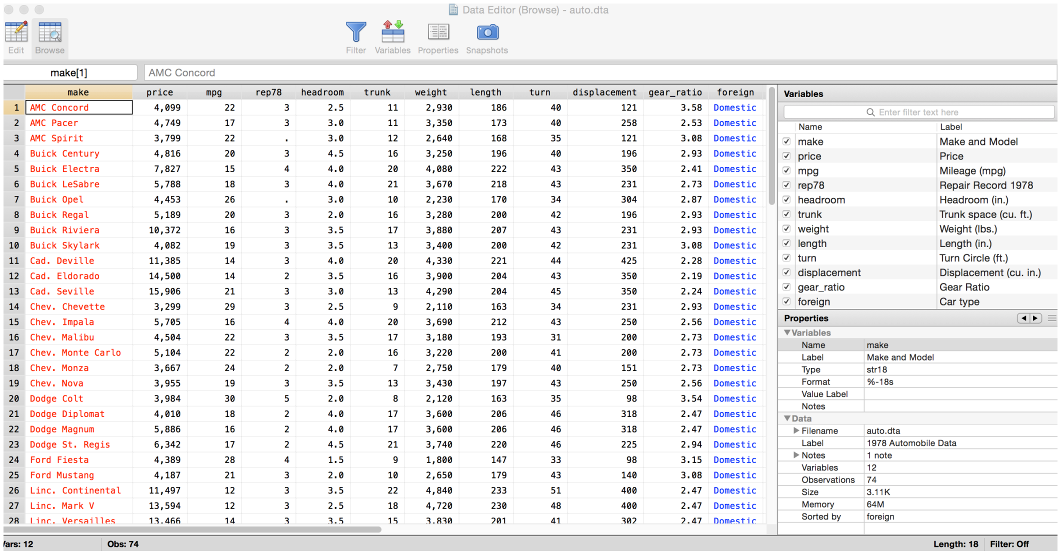
2.1 Looking at Variables

- ALWAYS start your data analysis by exploring your data.
- For this lecture, we will use Stata (what you'll use in most classes). Don't worry if you don't understand the code right now, it's just here so you can replicate the example if you want.
- The data set we are using is the 1978 Automobile Data provided by Stata.
- To view in the data in Stata, type the following into the command line:

```
sysuse auto
browse
```

- The following window should pop up:

[1] Contributor(s): Erin Fahle and Betsy Williams. If you find errors, please let us know so that we may correct them. Thanks!



Data Editor (Browse) - auto.dta

make[1] AMC Concord

	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio	foreign
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	40	121	3.58	Domestic
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	40	258	2.53	Domestic
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	35	121	3.08	Domestic
4	Buick Century	4,816	20	3	4.5	16	3,250	196	40	196	2.93	Domestic
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	43	350	2.41	Domestic
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	43	231	2.73	Domestic
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	34	304	2.87	Domestic
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	42	196	2.93	Domestic
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	43	231	2.93	Domestic
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	42	231	3.08	Domestic
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	44	425	2.28	Domestic
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	43	350	2.19	Domestic
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	45	350	2.24	Domestic
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	34	231	2.93	Domestic
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	43	250	2.56	Domestic
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	31	200	2.73	Domestic
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	41	200	2.73	Domestic
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	40	151	2.73	Domestic
19	Chev. Nova	3,955	19	3	3.5	13	3,430	197	43	250	2.56	Domestic
20	Dodge Colt	3,984	30	5	2.0	8	2,120	163	35	98	3.54	Domestic
21	Dodge Diplomat	4,010	18	2	4.0	17	3,600	206	46	318	2.47	Domestic
22	Dodge Magnum	5,886	16	2	4.0	17	3,600	206	46	318	2.47	Domestic
23	Dodge St. Regis	6,342	17	2	4.5	21	3,740	220	46	225	2.94	Domestic
24	Ford Fiesta	4,389	28	4	1.5	9	1,800	147	33	98	3.15	Domestic
25	Ford Mustang	4,187	21	3	2.0	10	2,650	179	43	140	3.08	Domestic
26	Linc. Continental	11,497	12	3	3.5	22	4,840	233	51	400	2.47	Domestic
27	Linc. Mark V	13,594	12	3	2.5	18	4,720	230	48	400	2.47	Domestic
28	Linc. Versailles	13,466	14	3	3.5	15	3,830	201	41	302	2.47	Domestic

Variables

Enter filter text here

	Name	Label
☑	make	Make and Model
☑	price	Price
☑	mpg	Mileage (mpg)
☑	rep78	Repair Record 1978
☑	headroom	Headroom (in.)
☑	trunk	Trunk space (cu. ft.)
☑	weight	Weight (lbs.)
☑	length	Length (in.)
☑	turn	Turn Circle (ft.)
☑	displacement	Displacement (cu. in.)
☑	gear_ratio	Gear Ratio
☑	foreign	Car type

Properties

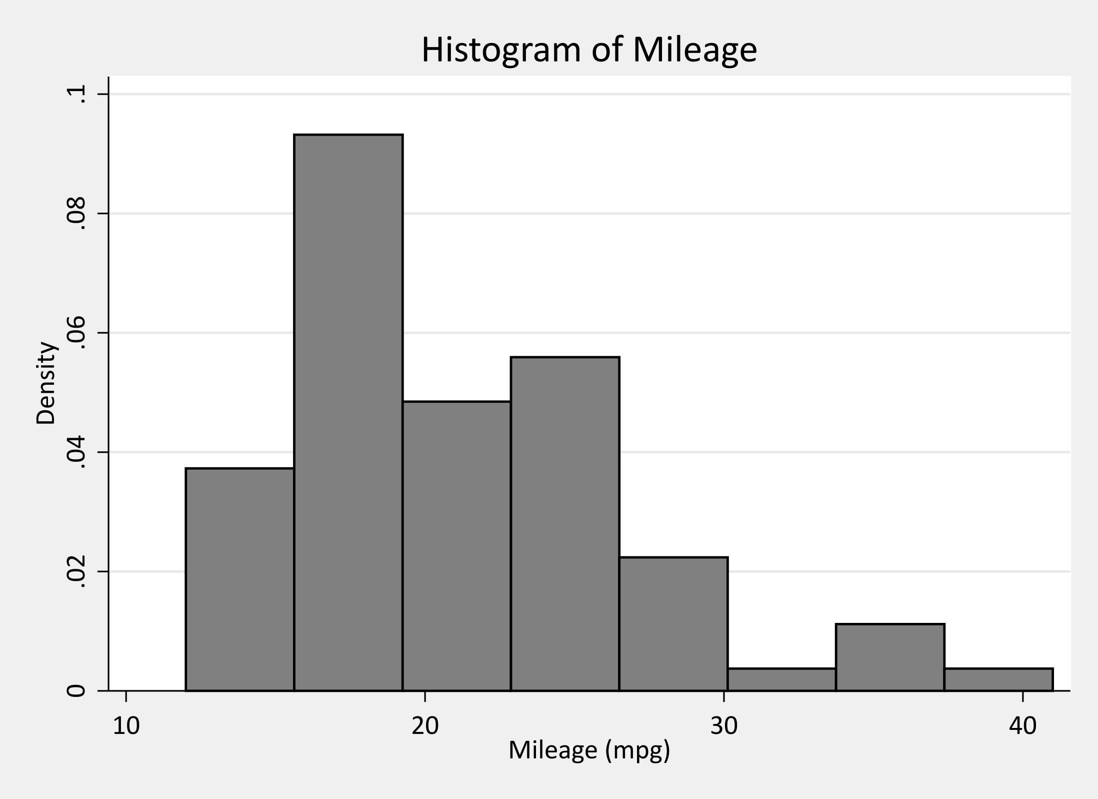
▼Variables	
Name	make
Label	Make and Model
Type	str18
Format	%-18s
Value Label	
Notes	
▼Data	
▶ Filename	auto.dta
Label	1978 Automobile Data
▶ Notes	1 note
Variables	12
Observations	74
Size	3.11K
Memory	64M
Sorted by	foreign

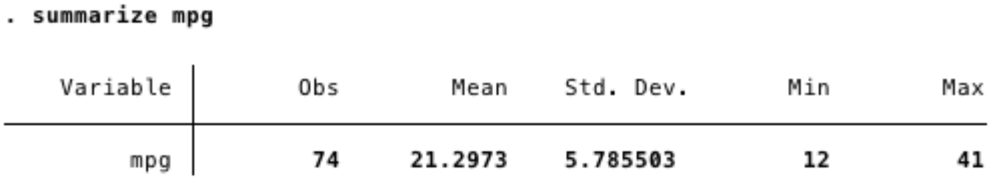
/ars: 12 Obs: 74 Length: 18 Filter: Off

- What you're looking at is a **data set** (a.k.a. **data table** or **data matrix**). The columns represent **variables** and the rows individual **cases**.
- Let's say we're interested in car mileage.
- To better understand mileage, you can plot histograms and calculate summary statistics for this variable.
  - A **histogram** provides a summary of the distribution of a **single** variable. The type of information we may look to gain from a histogram is the **shape** of the distribution of the variable (unimodal/bimodal, symmetric, skewed, and/or has outliers).
  - **Summary statistics** are numeric quantities that tell you about important aspects of the distribution of a variable, including the mean, standard deviation, number of observations, etc.
  - To make a histogram in Stata:

    hist mpg, title("Histogram of Mileage")
  - To calculate summary statistics:

    summarize mpg





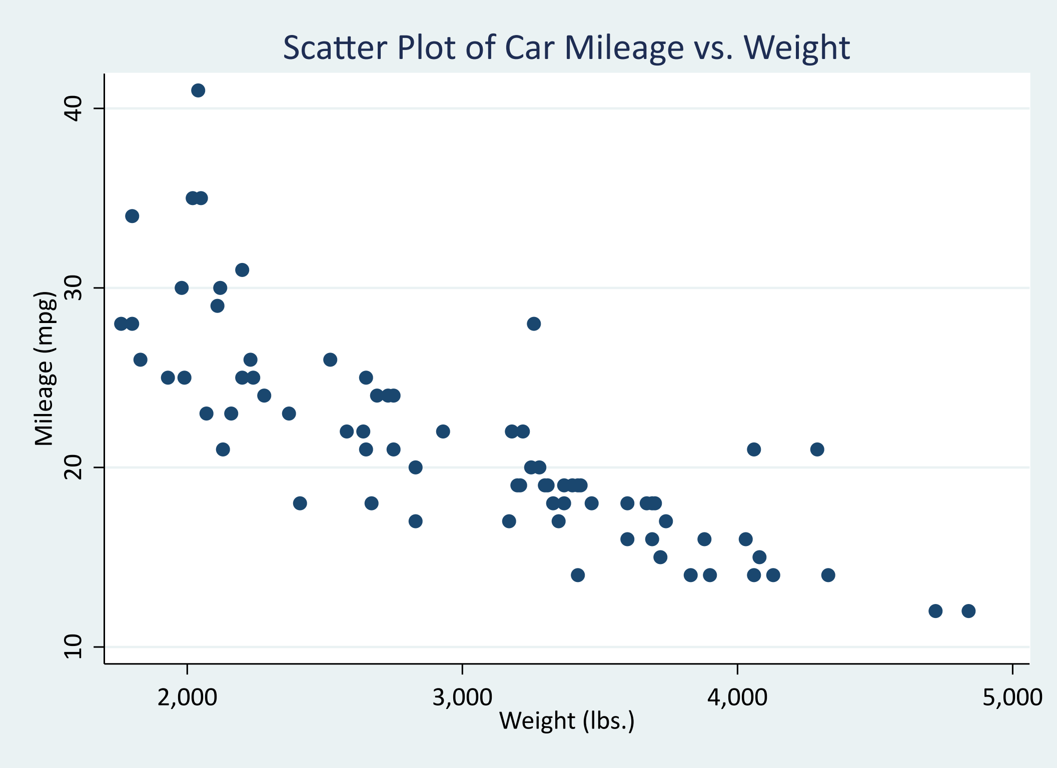
```
. summarize mpg

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         mpg |         74     21.2973    5.785503         12         41
```

- Describe the distribution of mileage per gallon (mpg). Is the distribution *approximately* normal?

2.2 Looking at Relationships

- To look at the relationship between two variables, we can use a **scatter plot**.
  - In a scatter plot, we plot the **dependent variable** against the **independent variable**.
  - Independent Variable: Typically denoted x, this is a variable that we think may predict our dependent variable.
  - Dependent Variable: Typically denoted y, this is the variable of interest, the one we are trying to understand better.
  - In this case, say we want to understand how weight affects mileage. We can plot mileage on the y-axis and weight on the x-axis to investigate this relationship.

    scatter mpg weight, title("Scatter Plot of Car Mileage vs. Weight")



- Describe the scatter plot:
  1. Does there appear to be a relationship between car mileage and weight?
  2. Is the relationship **positive** or **negative**, i.e. is the slope positive or negative? What does this mean?
  3. Draw a trend line through the points. What is the y-intercept of the line you drew? What is the slope? Keep these ballpark estimates in mind as we formally calculate it in the next section.

3 Linear Regression Using Stata

Now that we have seen that there appears to be a relationship between a car's weight and its mileage, let's model it using **Linear Regression**.

3.1 The Basics

- Linear regression finds the **best fit line** using a process called **Ordinary Least Squares** (OLS):

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

where:

β_0 is the **constant** (and Y-intercept)

β_1 is the **regression coefficient on X** (and slope)

ϵ is **error**[2]

[2] The error (or disturbance) is the variation we get in real-world data, and the error term, ϵ, represents this variation in the regression equation. We usually assume this error (or disturbance) is random.

- Why is it called **least squares**? It gets its name from the process we use to find the line. Specifically, we want to minimize:

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2}$$

  In words, we want to find the *least* sum of the *squares* of the differences between the observed values (y_i's) and the predicted values ($\beta_0 - \beta_1 x_i$).

  - NOTE: The differences between the observed and predicted values are called **residuals**. So the sum in the above equation is often called the **Sum of the Squared Residuals** (SSR or SSE).

- This is a calculus problem, but luckily someone already solved it! It can be shown that the least squares estimates are given by:

$$\hat{\beta_1} = \frac{Cov_{XY}}{Var_X} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

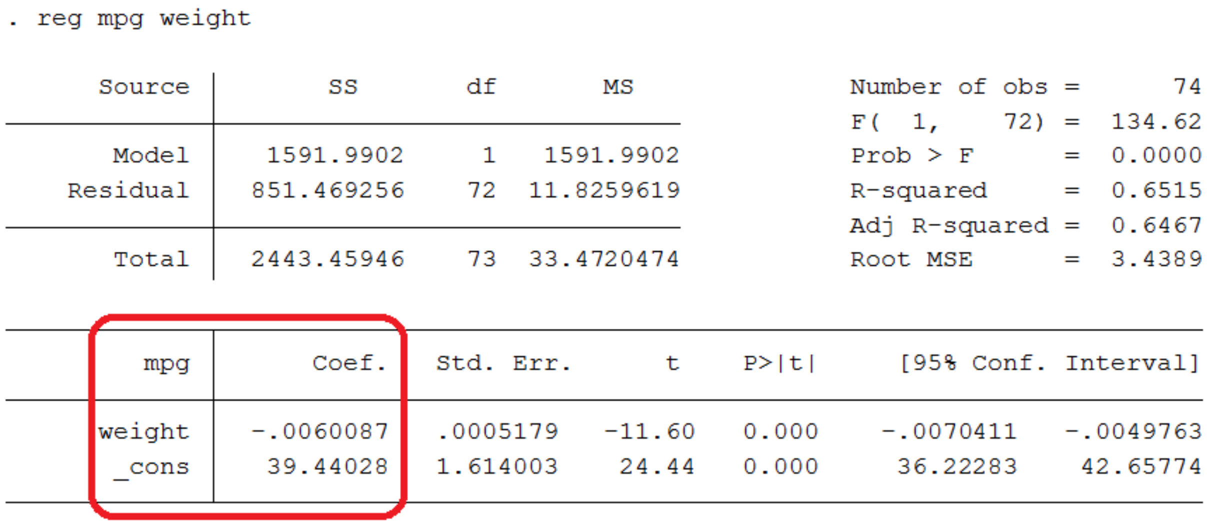
$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \tag{4}$$

- There are four key assumptions:
  - Assumption 1: The Linearity Assumption. The true relationship between the two variables is linear.
  - Assumption 2: Independence Assumption. The errors in the true underlying regression model are mutually independent.
  - Assumption 3: Equal Variance Assumption (Homoskedasticity). The variability of y should be about the same for all values of x.
  - Assumption 4: Normal Population Assumption. The errors are normally distributed.

3.2 Estimating a Linear Regression Line

- You'll most often use a statistical program to estimate regression lines. To estimate the regression line using Stata, type:

  regress mpg weight



```
. reg mpg weight

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  1,    72) =  134.62
       Model |   1591.9902     1   1591.9902           Prob > F      =  0.0000
    Residual |  851.469256    72  11.8259619           R-squared     =  0.6515
-------------+------------------------------           Adj R-squared =  0.6467
       Total |  2443.45946    73  33.4720474           Root MSE      =  3.4389

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0060087   .0005179   -11.60   0.000    -.0070411   -.0049763
       _cons |   39.44028   1.614003    24.44   0.000     36.22283    42.65774
------------------------------------------------------------------------------
```

- Using this output, we can write our regression equation as:

$$\widehat{Mileage} = \hat{\beta}_0 + \hat{\beta}_1 * Weight = 39.44 - .0060 * Weight$$

  - How do these estimated coefficients compare to your ballpark estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$?

3.3 Step 2: Looking at Predicted Values

- Let's use our estimated regression equation to predict a value of Y, given a value of X.

  1. Even if we have never observed a car that weighed exactly 3,000 pounds, we can predict how many miles per gallon it will get using the equation:

     $$\widehat{Mileage} = 39.44 - .0060 * Weight = 39.44 - .0060 * 3000 = 21.44$$

     Because 3000 is within the range of our Weight variable, this is called **interpolation**.

  2. What is the predicted mileage of a car that weighs 500 pounds?

     $$\widehat{Mileage} = 39.44 - .0060 * Weight = 39.44 - .0060 * 500 = 236.44$$

     We have not observed a car this light (i.e. this is outside the range of our Weight data), so this is **extrapolating**. Is this reasonable?

  3. Next, let's predict the mileage of a VW Rabbit, which is in the data. How does its estimated mileage, $\widehat{Y_{Rabbit}}$, compare to its actual mileage, Y_{Rabbit}?

     - A VW Rabbit weighs 1,930 lbs.
     - Using our regression equation, we predict it gets 27.86 mpg.
     - According to our data, it gets 25 mpg.
     - So we have overestimated the true mpg by 2.86 when using the regression line, i.e. there is error in our prediction of the VW Rabbit mpg.

- We can get Stata to predict each value of $\widehat{Mileage}$ for us.

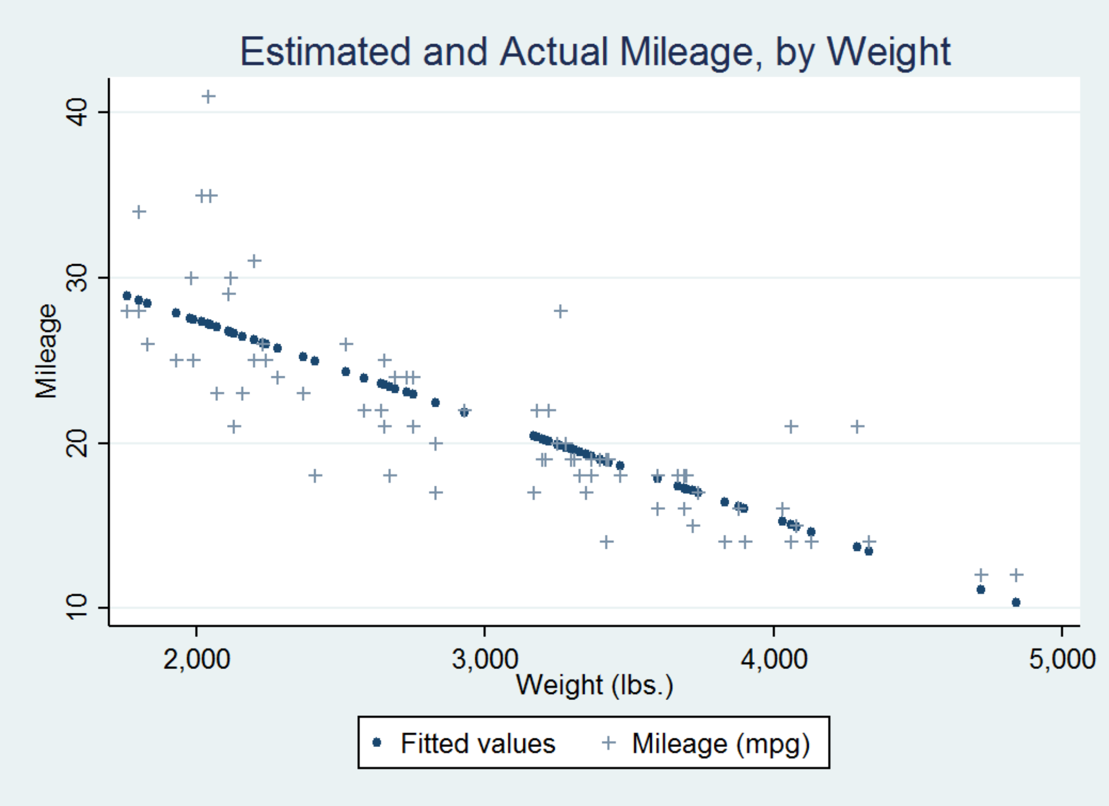
  predict mpghat

  twoway (scatter mpghat weight, msymbol(o)) ///

(scatter mpg weight, mcolor(emidblue) msymbol(+)) ///

, ytitle(Mileage) title("Estimated and Actual Mileage, by Weight")

- Here is a plot of both the predicted and actual mileage for each car.



3.4 Step 3: Look at the Residuals

- When we fit a line through many points, the line (no matter how well it is fit) cannot run through all the points (unless all points are colinear, but that usually never happens!).

- So, how can we say this function "fits the data," since there are clearly points that are not on the regression line we found?

- We need to look at the **residuals**. The residual is our best prediction of the unobservable error, ϵ, from our regression equation above.

  - The residual is calculated as: $r = Mileage - \widehat{Mileage}$
  - In Stata, type:

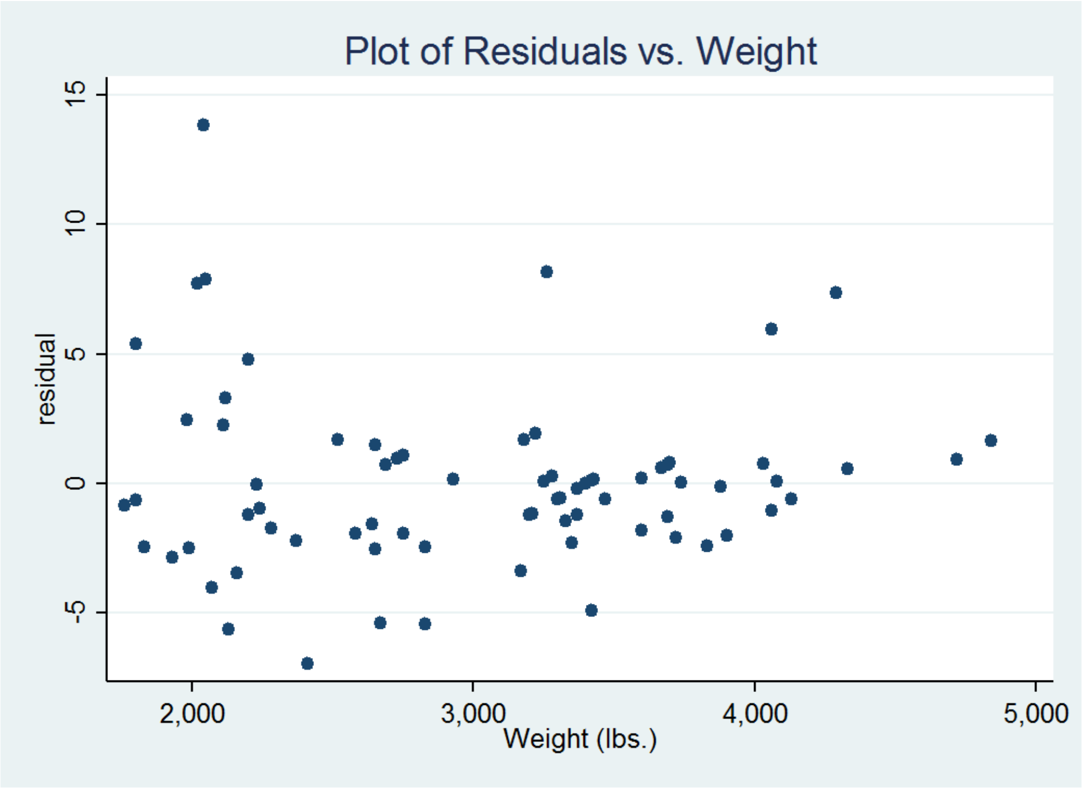
    generate residual = mpg - mpghat

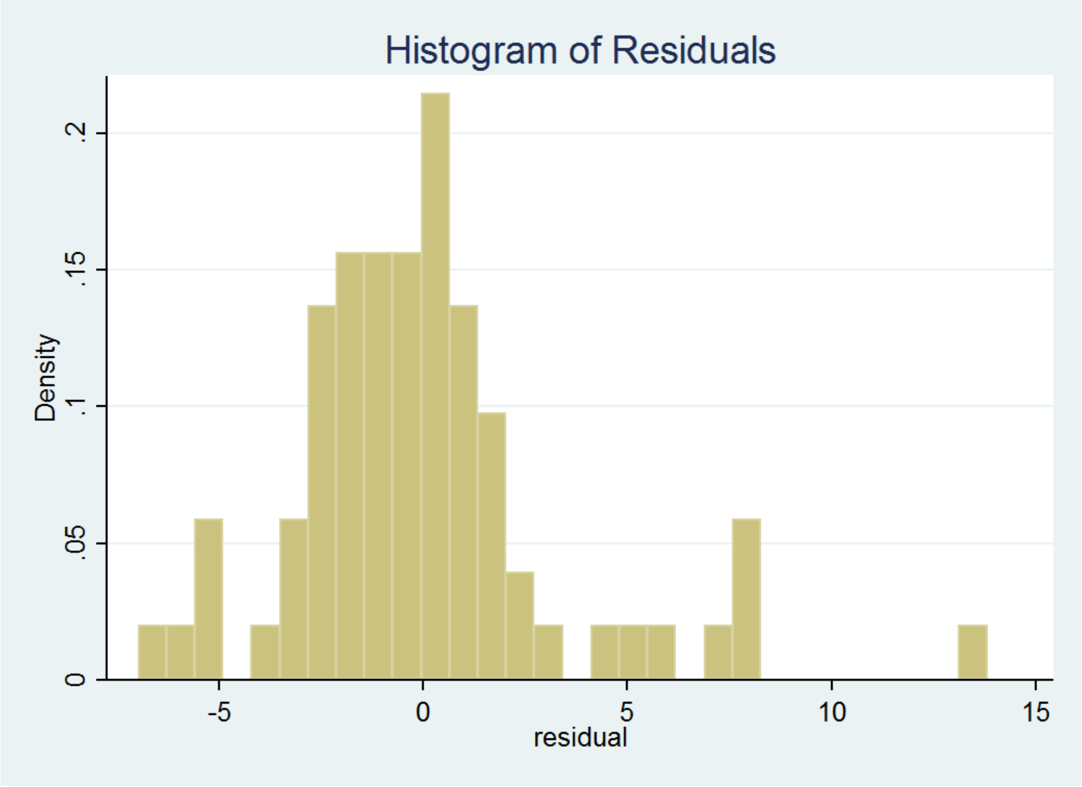
- We make a lot of assumptions about the **error** in linear regression that we need to check are satisfied by looking at the residuals.

  - Check Assumption 2: Plot your residuals against your x variable. They should look randomly scattered (uncorrelated/uniform) if our model fits right.
  - Check Assumption 3: Plot your residuals against your x variable. The spread around the line should be constant if the model fits right. Beware of "fan or funnel shapes".

- Check Assumption 4: Plot a histogram of the residuals. The distribution should look normal if they model fits right.
- In Stata, type:

  scatter residual weight, title("Plot of Residuals vs. Weight")
  hist residual, bin(30) title("Histogram of Residuals")





  - Are these three assumption met?
    - Assumption 2: The residuals appear to be uncorrelated with weight. If you drew a line through this plot, it would be nearly flat.
    - Assumption 3: It looks like there is more variance in the residuals for the lightest cars: this looks like it is maybe a case of **heteroskedasticity** – unequal variances! A violation of this assumption.
    - Assumption 4: The histogram appears normally distributed. Remember, we don't have that many observations so it will not be perfect!

- We can also calculate the sum of the squared residuals in Stata (REMEMBER: this is what we were trying to minimize!), choosing the variable names resid2 for the squared residual and sum_sq_resid for the sum of the squared residuals:

  ```
  gen resid2 = residual*residual
  egen sum_sq_resid = sum(resid2)
  display sum_sq_resid[1]
  ```

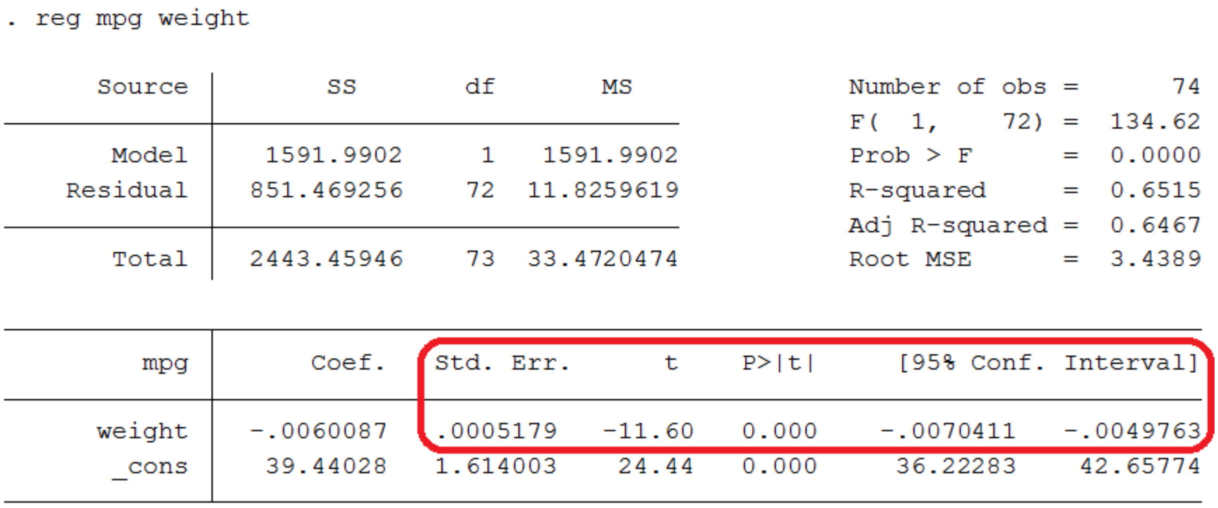
  SSR = 851.47. You can also find this in your Stata output, go back to it and check!

3.5 Hypothesis Tests and Confidence Intervals in Linear Regression

- We got an estimate of the slope in the regression equation, that relates the mileage per gallon of old cars to their weight: heavier cars get lower miles per gallon.

- But we want to know, is the regression coefficient on weight far enough from zero that we are 95% certain it is not zero? Are weight and mileage truly related in these cars?
  - We know there's variability: we could get different estimates of the slope depending on exactly which cars we include. (a.k.a. **sampling uncertainty**.) So there's variation in the slope estimate, and Stata helps us calculate that. (The calculations come from the rules for math with variances and covariances, expectations, and the Law of Large Numbers (LLN).)
  - In fact, the coefficient estimate is a kind of average, so the LLN applies and says that as the sample gets larger, the sample mean converges to the true mean. The Central Limit Theorem tells us the sample distribution of the coefficient is normal in a large enough sample. In a smaller sample, it is distributed as a related bell-curve, Student's t, with n-2 degrees of freedom.

- Let's look back at the regression output:



```
. reg mpg weight

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  1,     72) =  134.62
       Model |   1591.9902     1   1591.9902           Prob > F      =  0.0000
    Residual |  851.469256    72  11.8259619           R-squared     =  0.6515
-------------+------------------------------           Adj R-squared =  0.6467
       Total |  2443.45946    73  33.4720474           Root MSE      =  3.4389

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0060087   .0005179   -11.60   0.000    -.0070411   -.0049763
       _cons |   39.44028   1.614003    24.44   0.000     36.22283    42.65774
------------------------------------------------------------------------------
```

- We have an estimate, standard error, and we know the distribution should be roughly normal.

- The null hypothesis is that $\beta_1 = 0$.

- Our test statistic, t, is $\dfrac{\hat{\beta_1} - 0}{SE_{\hat{\beta_1}}}$. Stata calculated this to be -11.60. Do it by hand to confirm you get the same value.

  - Again, this calculation quantifies the difference between what we estimated our slope to be, $\hat{\beta_1}$, and its value according to the null hypothesis, 0.
  - Whether they are close enough to each other is relative to the standard error: SE provides a scale for how much variation we expect. A big SE means there is a lot of variability and a small SE that there is little variability.
  - This test statistic, t, shows us how many standard errors our estimate is from our hypothesized value.

- Stata looks up that test statistic in its distribution and reports how probable it is that we would see an estimate that extreme when the slope is truly 0, p-value $= P(|t_{\hat{\beta 1}}| > |t^*|\ given\ \beta_1 = 0)$. It also calculates the 95% confidence interval for us, also based on the test statistic and its distribution.

- Using Stata's calculations: can we reject the null hypothesis at the 5% level?

3.6 Other Estimates and Hypothesis Tests About the Regression

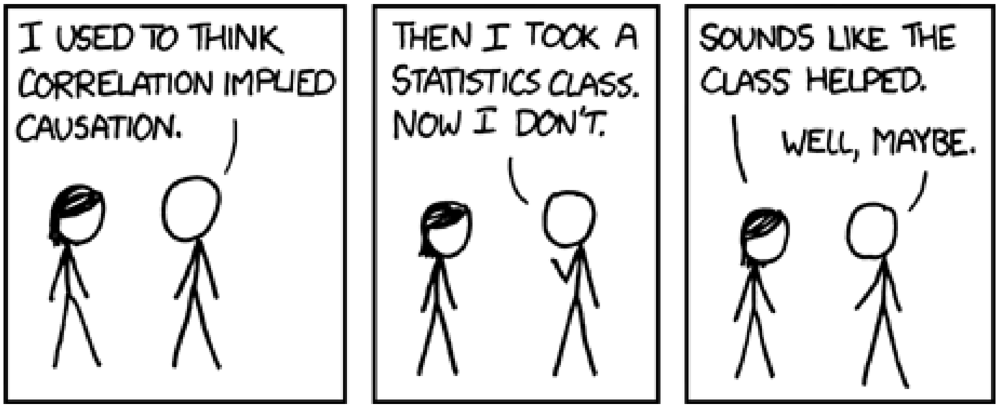
- There are also other things we could test about this regression.

  1. How good is the estimate of the intercept, $\hat{\beta_0}$?
     - This process is the same as the process for the slope coefficient. Try it!
     - Aside: do we want to use a one-sided test or a two-sided test to answer our question? (Think about your null and alternate hypotheses.) Which one do you think Stata chooses automatically?
  2. How good is the estimate of the entire model, taking both the slope and intercept estimates together?

    - This is a **joint hypothesis test**, that uses a different test statistic measured against a different distribution, the **F distribution**.
    - In the upper right corner, Stata reports the F statistic and its **p-value**. Again, this asks the probability of seeing a test statistic this extreme if the X variables we tested (Weight) actually had no linear relationship with Y (Mileage). Note: This becomes more interesting when there are multiple predictors because it is the simultaneous test that all coefficients (β_1, β_2, etc.) are equal to zero.
    - Look back at the regression output. Can you find F? What is the associated p-value?

- Stata's other regression output tells you how good the fit is, but there is no direct test to say that the values are "good enough."
  - The upper left shows a table of Sums of Squared Errors (SS) and Means of Squared Errors (MS) along with **degrees of freedom** (df).
    - The table separates what is explained by the model (**Explained Sum of Squares (ESS)**, labeled as Model SS or MSS in the output) and what remains unexplained (**Residual Sum of Squares (RSS)**).
    - Combined, ESS (or MSS) and RSS are the **Total Sum of Squares (TSS)**.
    - These SS and MS estimates are what the R^2 and Root MSE are calculated from.
  - The **Root MSE** is the square root of the average of the squared error estimates ("mean squared error"). This relates to the "Residual" row in the table, Stata labels the MSE as MSR. Root MSE is also called the **standard error of the regression**.
    - The closer it is to zero, the better the fit.
    - It is in units of Y, showing the size of the average residual. Here, it means the typical estimated value is off by 3.4 miles per gallon.
    - This is often referenced in psychology.
    - $MSE = MSR = SSR/(N - k - 1)$
    - Root MSE $= \sqrt{MSE}$
  - The R^2 and **Adjusted** R^2 show what percentage of the variation in the Y variable is explained by the right-hand side of the equation.
    - R^2 can range from 0 to 1.00.
    - We "adjust" R^2 when we have more than one predictor.
    - This is often referenced in social science.
    - $R^2 = MSS/TSS$
    - $R^2_{adj} = 1 - MSE/MST$

4 H/T to Causal vs. Correlational


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

From http://xkcd.com/552/ by Randall Munroe, Creative Commons Attribution-NonCommercial 2.5 License.

- Correlational: When X changes, Y tends to change. We observe a relationship, but we cannot prove why.
- Causal: Not only do we observe a relationship, but we see that X has a consistent **causal effect** on Y. But for X, Y would have a different value.
- How can we make our research (more) causal?
  - **Randomized Experiments**: randomize X and observe Y (the gold standard).
  - **Control variables** can fix omitted variable bias and make estimates more precise.
  - **Fixed effects** can compare observations within similar groups.[3]
  - **Regression discontinuity** shows a before-and-after story for a case where X changed, relative to a comparison group that did not have the same change in X.
  - **Instrumental Variables** finds one thing that has some causal effect on X, then uses only the variation induced by that **instrument** to predict the variation in Y.
  - **Propensity Score Matching**: Find people who are equally likely to have the same value of X and compare them to each other on their values of Y.
- The EDUC 255 sequence takes you through how to understand and implement these strategies to get estimates that are more causal.

[3] These are different than "fixed effects" in ANOVA.