Regression Lecture Notes[1]

Stanford GSE Math Camp 2018

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.
- We can explore each variable individually by making a **histogram** and describing its distribution as we discussed on Monday
  - Shape
  - Center
  - Spread
  - Unusual Features (Outliers, Gaps, etc.)

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.

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

- Describe the scatter plot:
  - **Form**: Does there appear to be a relationship between the x and y variable? Is it linear or non-linear (curved)?
  - **Direction**: Is the relationship **positive** or **negative**, i.e. is the slope positive or negative? What does this mean?
  - **Strength**: How strong does this relationship appear to be? Strong? Moderate? Weak? No relationship (cloud-like)?
  - **Unusual Features**: Does the scatter plot reveal a possible outlier? Are there clusters?

3 Pearson's Correlation Coefficient (r)

- Pearson's correlation coefficient (r) measures the strength of the linear association between two variables.
- The value of r can range from -1 to 1, where 1 is a perfectly positive linear relationship, 0 is a perfectly non-linear, and -1 is a perfectly negative linear relationship.
- To calculate r:
  - First, standardize each x and each y for every data point. This centers the scatterplot at the origin and scales the axes to standard deviation units.
    $$(z_x, z_y) = (\frac{x-\bar{x}}{s_x}, \frac{y-\bar{y}}{s_y})$$
  - Next, multiply the standardized score of each data point's x and y. Then, sum these products.
    $$\Sigma z_x z_y$$
  - Finally, divide by n-1.
    $$r = \frac{\Sigma z_x z_y}{n-1}$$
- This correlation coefficient is sensitive to outliers. A single outlier can have a large impact on the value.
- The correlation coefficient has no units.

4 Linear Regression Using A Stats Package

If we believe there to be a linear relationship between two variables, we can model it using **Linear Regression**.

4.1 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]

- 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:

$$\hat{\beta}_1 = \frac{Cov_{XY}}{Var_X} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

or

$$\hat{\beta}_1 = r\frac{s_y}{s_x} \tag{4}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \tag{5}$$

- 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.

4.2 Estimating a Linear Regression Line

- You'll most often use a statistical program to estimate regression lines. This is the output from Stata:

[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.

```
. 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:

$$\widehat{y} = \hat{\beta}_0 + \hat{\beta_1}(x)$$

$$\widehat{Mileage} = \hat{\beta}_0 + \hat{\beta}_1 * Weight$$

$$Mil\hat{e}age = 39.44 - .0060 * Weight$$

  - After finding our regression equation, we should interpret the values we get for the slope and the y-intercept.
    * **Slope:** For every one unit increase in x variable, the regression model **predicts** a value of slope (increase or decrease) in y variable.
    * **Y-Intercept:** When x variable is 0, the regression model predicts y variable to be value of y-intercept.

4.3 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?

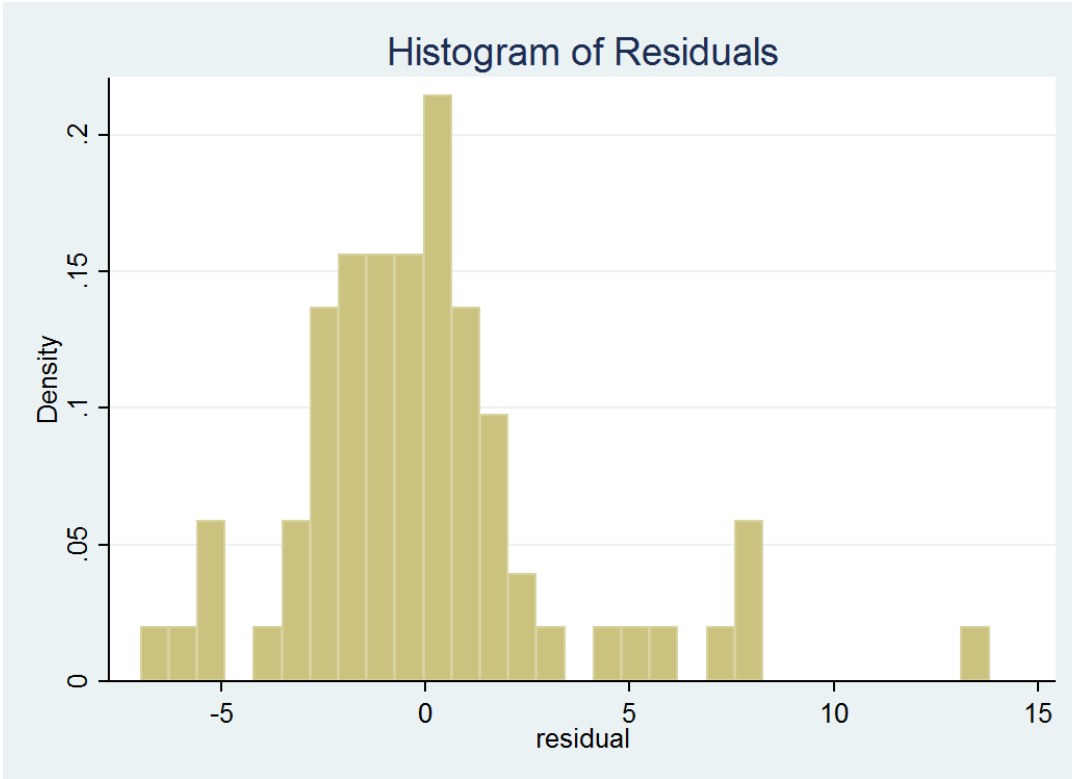
3. Next, let's predict the mileage of a VW Rabbit, which is in the data. How does its estimated mileage, $\widehat{Y_{Rabbit}}$, compare to its actual mileage, Y_{Rabbit}?
   - A VW Rabbit weighs 1,930 lbs.
   - Using our regression equation, we predict it gets 27.86 mpg.

- 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.

4.4 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: $e = y - \widehat{y}$
- 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")

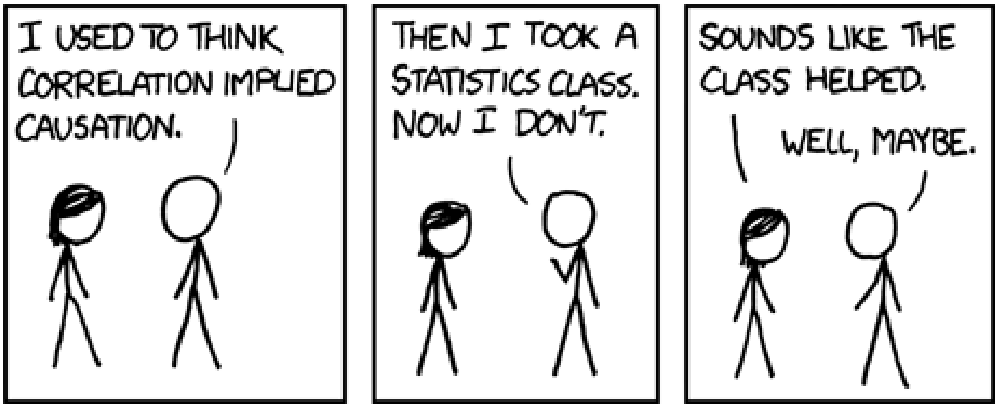


- Are these three assumption met?
  - Assumption 2: The residuals appear to be uncorrelated with weight. If you drew a line through this plot, it would be nearly flat.
  - Assumption 3: It looks like there is more variance in the residuals for the lightest cars: this looks like it is maybe a case of **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!

4.5 Coefficient of Determination, (R^2)

- 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.
  - $R^2 = r^2$

5 H/T to Causal vs. Correlational



- 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.