Chapter 5

Linear Regression

This ActiveStats document contains a set of activities for Introduction to Statistics, MA 207 at Carroll College. This is a non-calculus based statistics class which serves many majors on campus. This document is intended for the classroom teacher to support students in active engagement with statistics on a daily basis. This document is not designed to be given to students as is. Rather, it is a teacher resource.

The activity set is designed to work alongside the OpenIntro *Introductory Statistics with Randomization and Simulation* textbook by Diez, Barr, and Cetinkaya-Rundel. The chapters in ActiveStats are numbered to align with OpenIntro, though the subsections may differ. OpenIntro is an open source curriculum with accompanying data sets. OpenIntro is the textbook resource to direct students to for out-of-class reading assignments and review. We also use the Cartoon Guide to Statistics as a supplement for assigned reading.

Data sets for ActiveStats can be found at mathquest.carroll.edu/activestats/data/ or on the class Moodle page.

5.1 Introduction to Linear Regression

Reading Assignment 5.1. Read Section 1.6.1 (p20-21) and Section 5.1 (p219-227) in OpenIntro Statistics.

Preview Activity 5.2. Understanding Scatterplots

1. According to Figure 1.16 on page 20 of Diez, et al, what can we say about the relationship between the number of characters and the number of line breaks? On average, does it seem like when *num_char* increases, *line_breaks* increases or decreases? Is this true for every email in the email50 collection?

2. Do your answers to the previous questions, which were based on the scatterplot only, agree with your intuition about the *num_char* and *line_break* variables? Why or why not?

3. Suppose that I have one email with 20,000 characters and another with 30,000 characters. If you had to guess which had more line breaks based on Figure 1.16, which would you pick? About how many *more* line breaks would it have?

4. Does your answer to the question above agree with your intuition? Why or why not?

5. Suppose that I have an email with no characters (boring email, no?). Based on Figure 1.16, how many line breaks should I expect in this email? Does this agree with your intuition? Why or why not?

6. In Figure 1.16, *num_char* appears on the x-axis, and *line_break* appears on the y-axis. Mathematically, it would be just as easy to switch these (to put *num_char* on the y-axis, and *line_break* on the x-axis). Explain the significance of the choice made in the book about variable placement (Hint: you may want to use the terms **explanatory variable** and **response variable**).

▲

Activity 5.3. Used Honda Civics: Download the Used_Honda_Civic.xls data. This data set contains ages and prices for 90 Honda Civics. We are interested in assessing *whether the age of a used car can be used to predict price*.

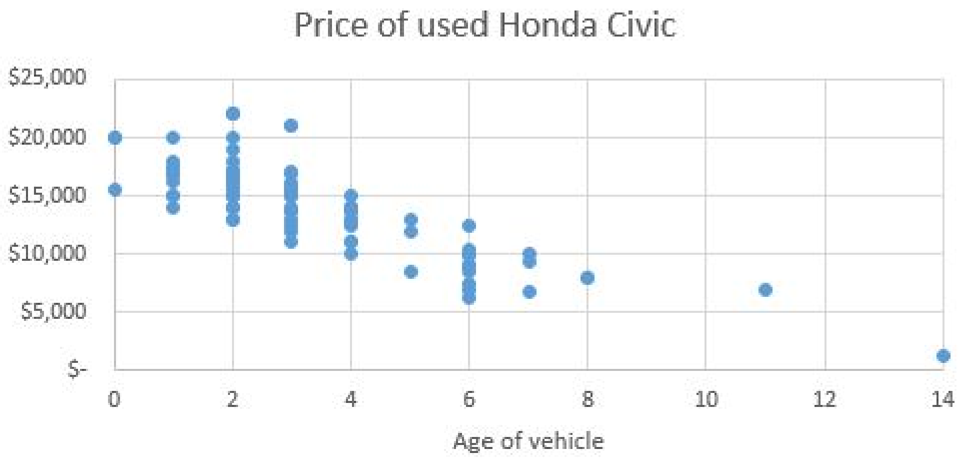
Part I: Initial Explorations.

1. In this analysis, what is the explanatory (or independent) variable? What is the response (or dependent) variable? Would reversing the variables make sense?

2. Use a pivot table to find the average age and average price of Honda Civics

3. Before creating a scatterplot, comment on whether you think there is a relationship between these values.

4. Now create a scatterplot of the data with age on x-axis and price on the y-axis. Does the scatterplot agree with your answer to the previous question? Explain.



5. Where does the point $(\overline{x}, \overline{y})$ land in relation to the other data points on your scatterplot. Does this surprise you? Explain.

6. Do you see a positive relationship (when x increases, usually y also increases), a negative relationship (when x increases, usually y decreases), a nonlinear relationship (depending on the value of x, increases x affect y differently), or no relationship?

7. Is there any chance that there is no true relationship between these variables. That is, is there a chance that it's just by random chance alone that the scatterplot looks like this? (Hint: The answer is yes – but how likely or unlikely do you think this is?)

Technique 5.4. To relate two variables where one variable is presumed to depend *linearly* on another we often find a *best fit line* of the form

$$y = mx + b,$$

where x is the explanatory (or independent) variable and y is the response (or dependent) variable.

Formulas for m and b can be found in Chapter 5 of Diez, et al. However, in our class, we will be using the Excel commands $=SLOPE(),$ and $=INTERCEPT()$ to find these values.

Handy facts about regression:

- Regression allows for interpolation and extrapolation (but extrapolation is dangerous!)
- Regression is one of the most used (and abused!) statistical tools.
- The point $(\overline{x}, \overline{y})$ is always on the best fit line.
- The slope of the line is built from the standard deviations in the x and y directions.

Part II: Finding the best-fit line.

1. Write down the equation of the best-fit line.

2. Right click on any data point in your scatterplot and select 'Add Trend Line' and then 'Display Equation.' You should see the same equation that you wrote down before displayed on your chart, along with a plot of the line among your data points.

3. Notice that many of your data points do *not* fall on this best-fit line. This begs the question: What good is this line then? Well, ... what good is it, do you think?

Part III: Understanding Slope and Intercept.

1. Does your slope coefficient agree with your answer to Question **??** above?
2. How does the slope affect the trend line we just drew?

3. How is the slope you calculated related to used Honda Civics?
4. How many lines with this slope could we have drawn?
5. How did Excel choose which of these lines to draw?
6. According to the best-fit line, what is the expected price of a brand new Honda Civic?
7. Do you think this number is accurate? Why or why not?
8. What does this number have to do with your best-fit line?

Next, we want to assess the *strength of the relationship between our variables*

Definition 5.5. The correlation coefficient R, describes the strength of the relationship between two variables.

- Negative values of R suggest that the relationship is negative, while positive values of R suggest that the relationship will be positive.
- The stronger the relationship, or 'fit' of these variables, the further R will be away from zero.
- R^2 is the percent of the variation in y explained by x.
- A formula for R can be found in Chapter 5 of Diez, et al. However, in our class we will use the Excel command $= CORREL()$.

Concept check: If the strength of a linear relationship is demonstrated by the distance from R to 0, what does it mean when R is really close to zero?

Part IV: Understanding R and R^2. We want to assess the strength of the relationship between our variables.

1. Before calculating R, make a guess about what you think it will be in the case of age of Used Honda Civics against their prices. (Remember that -1 and 1 represent perfect fits, and 0 represents no relationship at all)

2. Now find the correct value for R. Does it line up with your guess?

3. R^2 tells us the percent of variation in car prices that can be explained by the age of the car. If $R^2 = 0.4$, then 40% of the price can be explained by the car age, but the other 60% is due to other factors. What other factors would likely affect the price of a used Honda Civic? (Note: R^2 is higher than 0.4 for this data set.)

4. You can find R^2 by squaring R or by adding it to your plot using the Trendline options.

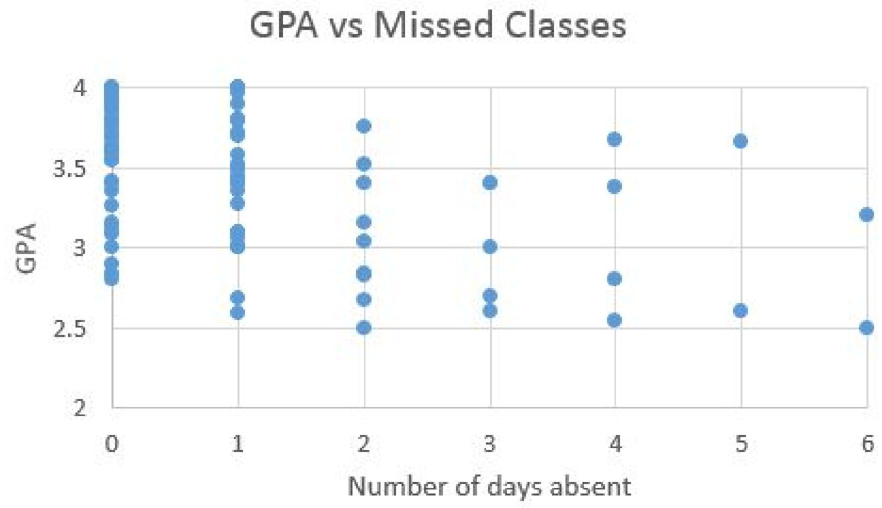
Part V: Practice. Answer the following questions:

1. Assume that we have an 8 year old Honda Civic. According to our linear regression, what is the approximate price if we were to sell it?

2. According to our linear regression, what is the approximate price of an 18 year old Honda Civic?
3. If a used car dealer sold a Honda Civic for $10,000, what was the approximate age of the car according to our linear regression?
4. What percent of the variation in price is explained by the age of the Honda Civic?

▲

Activity 5.6. GPA vs. Missing Class:
A high school athletic director is worried that his athletes are missing too much class. In particular, he is curious if there is a relationship between the GPA of his student athletes and the number of classes missed during a semester for sports-related reasons. Data for this task can be found in the file GPAMissingClasses.xlsx.



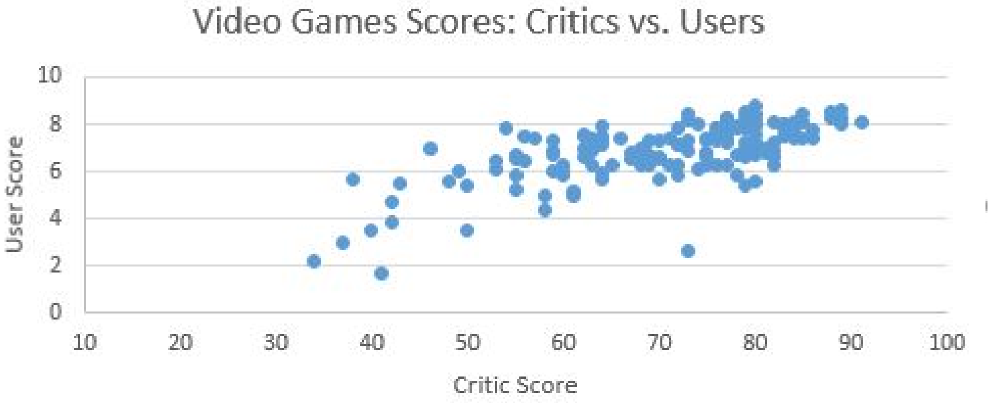
1. Create a linear regression model for this data.
2. Classify the relationship as strong / weak, positive / negative, linear / nonlinear
3. How much variation in GPA is explained by the number of days missed?
4. According to the best-fit line, what is the expected GPA for a student who misses 3 days for sports? Does this seem reasonable?
5. According to the best-fit line, what is the expected GPA for a student who misses 7 days for sports? Does this seem reasonable?

▲

Activity 5.7. Video Game data set
Open the VideoGamesX360PS3.xlxs data set. What numeric categories might be related to each other? Which data could be used to predict other data?

1. Is there a linear relationship between how critics score the games and how users score the games? (Let x = critic score and y = user score). *For this question, we could swap the independent and dependent variables and still get reasonable results.*

   Start by creating a visual model of the data. You may notice that user scores are much lower than critic scores. Why is that?



2. How strong is the relationship between critic scores and user scores? *Describe it in both words and numbers* Is it stronger or weaker than you anticipated? What does that mean for the video game industry?

3. Making predictions:

   (a) For a game with a critic score of 80, what would you predict for the user score?

   (b) For a game with a user score of 5, what would you predict for the critic score?

   (c) Your boss at Video Games R Us wants to know if you can predict the user score for a video game where the critic gave a score of 15. What would you say?

4. (Data Clearning Challenge) Sometimes researchers ignore data if the data doesn't meet certain requirements. Let's choose to ignore data where we don't have at least 10 users providing scores and at least 10 critics providing scores. (This information is provided in CriticCount and UserCount). Create a new scatterplot of this limited data and determine if removing those data points had a strong impact on either the trendline or the R^2 value.

Going further: Do user scores relate to sales? Do critic scores relate to sales? Do sales in one region predict sales in another?

▲

5.2 Regression: Outliers, Residuals, and Interpreting Regression Coefficients

Reading Assignment 5.8. Read Section 5.2 and 5.3 in the OpenIntro Statistics book.

Preview Activity 5.9. Outliers and Regression

Navigate to https://www.geogebra.org/m/W9qJrwmb to answer the following questions:

1. If you have a strong positive correlation, how can one outlier influence the correlation coefficient?
2. Can an outlier switch a correlation from positive to negative?
3. What do you do when you spot an outlier in your data set?

▲

Recall the following vocabulary, which we will be using today:

Definition 5.10. **Outliers** are points which do not follow the trend exhibited by the rest of your data.

Leverage points are points whose x-values are far away from most of the other x-values in your data.

For today's activities we will be using data found in the Framingham Heart Study:

> The Framingham Heart Study, a project of the National Heart, Lung, and Blood Institute, is an ongoing longitudinal cohort study focused on citizens of Framingham, Massachusetts. This study began in 1948, and continues today in collaboration with Boston University. Researchers follow cohorts of people to learn more about the epidemiology of cardiovascular disease. Each cohorts' population is comprised of the offspring of each prior cohort, and researchers use this information to observe trends and familial links to heart disease.
>
> This landmark research has strengthened the body of evidence supporting the link between lifestyle, environment, inheritability, and heart disease. Important findings of this research are now known risk factors for heart disease, of which high blood pressure, high cholesterol, smoking, obesity, diabetes, and a sedentary lifestyle are currently the most prevalent (History of the Framingham Heart Study, 2017). Another important product of this study is the Framingham Risk Score, which indicates the risk an individual has of having a cardiovascular event within the next 10 years. This tool is used worldwide to stratify risk categories. It is estimated that there are over 1,000 publications from this study, with more being added every year.

Taken from The History of the Framingham Heart Study. (2017).
https://www.framinghamheartstudy.org/about-fhs/history.php

Download the data set *Framinghamdata.csv* from the ActiveStats website.

Activity 5.11. Outliers in Linear Regression: For this activity we are going to focus on the variables *BMI* and *DBP*. We are asking the question "Is body mass index a good predictor of diastolic blood pressure?"

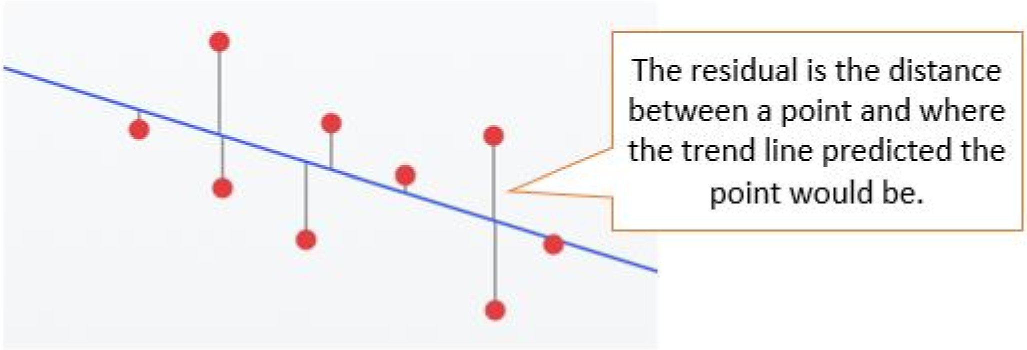
1. Create a data visualization to explore this question before doing any calculations.
2. Do you notice any outliers or leverage points? Pick three data points which seem like possibly problematic outliers or leverage points.
3. Add a trend line to your plot, and write down both the equation of your best-fit line and the correlation coefficient.
   - Is the relationship between these variables strong or weak? Positive or negative?
   - What is the real-world meaning of the slope coefficient?
   - What is the real world meaning of R^2 here?
4. Now try eliminating each of your chosen problematic data points (outliers and leverage points), one-at-a-time. Each time you do this, record the new trend line equation and new correlation coefficient. After eliminating an outlier from your data and finding the trend line, add it back to the data set before you eliminate the next one.

   Which outlier or leverage point had the greatest influence on your linear regression? Can you explain why? Which one had the least impact on your regression? Can you explain why? Would you qualify any of these points as influence points?

▲

Definition 5.12. The **residual** for each data point is its vertical distance from trend line.

- Positive residuals come from points above the trend line, and negative residuals from points below the trend line.


The residual is the distance between a point and where the trend line predicted the point would be.

Activity 5.13. Residuals in Linear Regression. For this activity we will continue to focus on the question "Is body mass index a good predictor of diastolic blood pressure?"

1. Working with a partner, try to invent a method in Excel which will calculate the residual for each of your data points. When you have a method you think will work, submit a brief description to the online survey, so that we can discuss your ideas.
2. Find the residuals of each data point.
3. Create a scatterplot which has BMI on the x-axis and your residuals on the y-axis.
4. Describe the plot you've just created.

▲

Activity 5.14. Interpreting Regression + Relationship Strength and Significance:

We wish to study relationships between three pairs of variables (one pair which we've already explored a bit, and two new pairs).

(A) BMI (body mass index) and DBP (diastolic blood pressure)

(B) HRTRT (heart rate) and BMI (body mass index)

(C) DBP (diastolic blood preassure) and SBP (systolic blood pressure)

1. For each of the pairs listed above
   - Generate a scatterplot with the first variable on the x-axis and the second variable on the y-axis. Comment on whether you see a noticeable relationship, and whether the relationship is positive or negative.
   - Find the slope, intercept, and correlation coefficient.
   - Write down (in plain English sentences), what each of these numbers means, in terms of the variables in question.

2. Which of the three relationships above is the strongest? The weakest?

3. Do you think all three represent true underlying relationships between variables, or do you think that one or more might just be due to random chance?
4. Do you have any ideas about how we could answer the previous question using statistics?

▲

5.3 Inference with Linear Regression

Reading Assignment 5.15. Read Section 5.4 in the OpenIntro Statistics book.

Preview Activity 5.16. When is Linear Regression Appropriate?

Important Assumptions 5.17. The following assumptions should be satisfied in order to assume that a best-fit regression line is a *valid* way to model a data set.

- The data should appear linear
- The variability around the line needs to remain roughly constant
- The residuals need to be nearly normal
- The observations must be independent

1. Draw a picture of a scatter plot that violates the first bullet
2. Draw a picture of a scatter plot that violates the second bulletl
3. Draw a picture of a scatter plot that violates the third bulletl
4. Draw a picture of a scatter plot that violates the fourth bulletl

▲

Activity 5.18. In small groups, students should discuss, agree on, and draw the scatterplots they created in the preview activity. Dissect a few to show whether they do or do not accomplish the task they were meant to.
▲

Activity 5.19. Residual plots provide a visual to focus on how far data points are from where the regression line predicts them. By plotting just the distance from the regression line, we are able to look for patterns in the plots which might suggest whether or not a linear relationship is an appropriate.

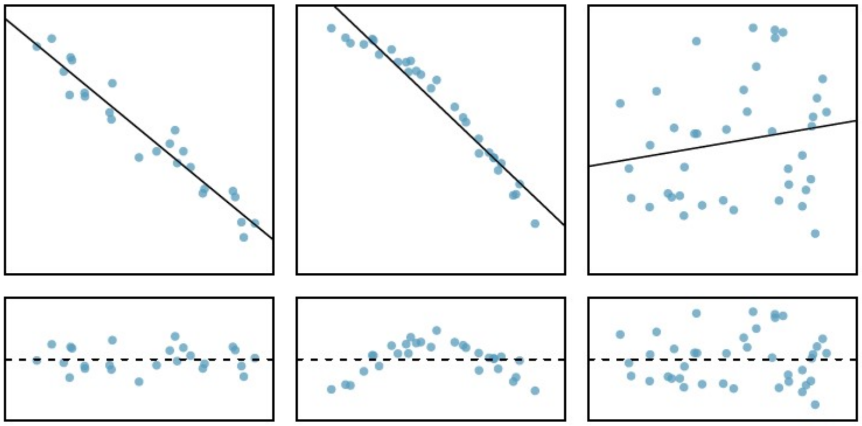


Image from OpenIntro Statistics book.

▲

Activity 5.20. Used Car data
Part I: Checking Assumptions. Open the Used Car data to assess whether linear regression is appropriate here. Create a scatterplot of car age vs. car price (or use one you previously created).

1. Does the data appear to be roughly linear?
2. Calculate residuals.
3. Create a residual plot in Excel to check the second bullet (plot age vs residual)
4. Copy the residuals into TinkerPlots to create a histogram.
5. Do the Honda Civic data satisfy the third bullet?
6. Do you notice any troubling outliers or leverage points? If so, how should we handle them?

Next, we want to ask the question: "Is the relationship between age of used Honda Civics and their prices statistically significant"? But first, we need the following tools.

Technique 5.21. Statistical Inference for the Slope:
If none of the regression assumptions listed in Important Assumptions **??** are violated, we can use a t-distribution (with $df = n-2$) to analyze the slope.

Null Hypothesis: *The slope is zero*
Alternative Hypothesis: *The slope is not zero*

The point estimate is $\hat{m}$, the estimated slope, and the standard error is given by

$$Std\ Error = \frac{s_y}{s_x}\sqrt{\frac{1-R^2}{n-2}}$$

Part II: Testing for significance.

1. What does the null hypothesis mean in the context of this linear regression?
2. Find the relevant standard error for this hypothesis test, and determine whether you will reject the null hypothesis.
3. If you reject the null hypothesis, you should provide a confidence interval for the slope. State it as part of a complete sentence describing the relationship between your variables. Finish your analysis with one or more complete sentences to answer our guiding question: "Is the relationship between age of used Honda Civics and their prices statistically significant"?

▲

Activity 5.22. Now let's check whether regression is appropriate for the GPA vs. Number of Classes missed example, and whether this relationship is significant.

1. Check to see if linear regression is appropriate for the GPA vs Missed Classes data. If not, what assumptions are violated?
2. Check to see if there are any outliers or leverage points. How will this affect our regression?
3. What are the null and alternative hypotheses in this example?
4. Find the point estimate and the standard error. Will you reject the null hypothesis?
5. Finish your analysis with one or more complete sentences to assess statistical significance and, in the case of significance, give a relevant confidence interval.

▲

Activity 5.23. Now let's check whether regression is appropriate for the DBP vs BMI example, found in the Framingham data set, and whether this relationship is significant.

1. Check to see if linear regression is appropriate for the DBP vs BMI data. If not, what assumptions are violated?
2. Check to see if there are any outliers or leverage points. How will this affect our regression?
3. What are the null and alternative hypotheses in this example?
4. Find the point estimate and the standard error. Will you reject the null hypothesis?
5. Finish your analysis with one or more complete sentences to assess statistical significance and, in the case of significance, give a relevant confidence interval.

▲

5.4 Regression Lab

Activity 5.24. Linear Regression Lab

In this lab, students look at the KidIQ.csv data set which comes from the National Longitudinal Survey of Youth. Students explore the relationship between mothers' IQ score and childrens' IQ score, as well as mothers' age and childrens' IQ score. Students find and interpret trend lines and correlation coefficients. They also conduct hypothesis tests for the significance of the slope.

▲

5.5 Summary

Summary 5.25. Student learning outcomes from Chapter 5

1. Students are able to create scatterplots of bivariate data and recognize situations that are appropriate for linear regression.

2. Students are able to create trend lines and appropriatetly interpret the meaning of the slope and y-intercept as they relate to the context of the problem.

3. Students understand the meaning of the correlation ccoefficient R and R^2.

4. Students recognize outliers and leverage points in data. Students are able to create and interpret residual plots to determine if a linear trend line is appropriate.

5. Students are able to conduct hypothesis tests on the signficance of m, the slope of the regression line.