MTH 4230 Spring 2021
Module 3 Notes and Exercises

Simple Linear Regression with One Predictor and Normal Error

A simple linear regression **model equation** for the value of a response variable Y using a single predictor variable, X, is given by

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where ϵ is a normally distributed random variable with mean 0 and variance σ^2.

Model Parameters The model equation is a three-parameter family of statistical models. Each choice of β_0, β_1, σ^2, corresponds to a specific member of the family. The model is also called a *first order, simple, linear model with normal error*. The parameters β_0 and β_1 are also called the *regression coefficients*. Parameters can be interpreted as follows.

- The slope, β_1, of the least squares regression line for the bivariate population: this parameter can be interpreted as the average change in the response with respect to a unit increase in the predictor.
- The y-intercept (intercept with the line $x = 0$), β_0, of the least squares regression line for the bivariate population: this parameter might not have a direct practical interpretation if the predictor value $x = 0$ is not within the range of the experimental data.
- The variance, σ^2, of the (normally distributed) residual errors associated with the least squares regression line for the bivariate population: this parameter yields the practical interpretation that 68/95/99% of possible values of Y given that $x = x^*$ should be within one/two/three standard deviations, σ, of $\beta_0 + \beta_1 x^*$.

Sample Statistics A bivariate simple random sample from an underlying population is denoted $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. The population notation is the same except that N replaces n.

In observational studies, researchers may not be able to control predictor values, and measurement error of predictors is a serious concern. In such cases, analytic methods for determining the sensitive of the model to measurement error have been developed.

In controlled experiments, predictor variable values can be adjusted to the researchers specifications. Typically, predictor variable values equally partitioned over the domain of interest are preferable, but this may depend on the goals of the analysis.

The following sample statistics are used in deriving inferential methods described below.

$$\begin{aligned} S_{xx} &= \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} \\ SST &= S_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} \\ S_{xy} &= \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n} \\ b_1 &= \frac{S_{xy}}{S_{xx}} \end{aligned}$$

$$
\begin{aligned}
b_0 &= \overline{y} - b_1\overline{x} \\
s^2 &= \frac{SSE}{n-2} \\
SSE &= \Sigma(y_i - \hat{y}_i)^2 = \Sigma y_i^2 - b_0\Sigma y_i - b_1\Sigma x_i y_i \\
r^2 &= 1 - \frac{SSE}{SST} \\
s_{b_1} &= \frac{s}{\sqrt{S_{xx}}} \\
s_{\hat{y}} &= s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}
\end{aligned}
$$

Inference: Point Estimates A **point estimate** of a population parameter is, in general, any statistic calculated from a sample and used to estimate a population parameter. A point estimate does not need to be a good estimator of a population parameter to fit the definition, and many point estimates of the same population parameter can be constructed. What it means to be a "good" point estimate depends on the properties desired by the statistician.

The specific point estimate a statistician employs satisfies specific criteria defined in theoretical statistics. For example, statistics that are "on target" and the most "reliably close" to their parameter are called **unbiased, minimal variance estimators**. When the population distribution is assumed to have a parametrized form (e.g. μ and σ for a normal distribution), a value of a parameter that maximizes the likelihood of the observed data is called a **maximum likelihood estimator**.

The following point estimates for simple linear regression parameters are developed in statistical theory.

- A point estimate for β_1 is given by $b_1 = \frac{S_{xy}}{S_{xx}}$.
- A point estimate for β_0 is given by $b_0 = \overline{y} - b_1\overline{x}$
- A point estimate for the mean value of the response variable given that the predictor variable equals x^*, denoted $\mu_{Y|x^*}$, is given by $b_0 + b_1x^*$.
- A point estimate of σ^2 is given by $s^2 = \frac{SSE}{n-2}$; hence, a point estimate of σ is given by $s = \sqrt{\frac{SSE}{n-2}}$

A theorem called the Gauss-Markov theorem establishes b_1 and b_0 have minimum variance among all *unbiased linear estimators*. The estimator s^2 can be identified as a *restricted maximum likelihood estimator*.

We cover further inference in Module 4, including confidence interval estimates and hypothesis testing techniques for making decisions about parameter values.

Model Diagnostics When using a mathematical or statistical model, it is important to perform diagnostic tests to determine if the model is appropriate. The objective of **statistical model diagnostics** is to determine if the data under consideration fits the model assumptions. For the simple linear regression model with one predictor and normally distributed error, the following diagnostic techniques are routine.

Real World Considerations In many models, sometimes obvious real world considerations suggest that simple linear regression is not an appropriate model for a bivariate data set. For example, if bone growth as a function of age of a subject is of interest to a researcher studying human growth, it is obvious that bone length is not a linear function of age over a normal human lifetime.

On the other hand, if we restrict the domain of predictor variable, it is often safe to assume that a bivariate data set is well approximated by a simple regression model over a constrained range of predictor values. For example, bone growth as a function of age of a subject could be approximated as linear over a constrained range of age values (say, ages 2-3), and the slope of the regression model could be interpreted an approximation of rate of growth for two-year-olds.

Before data collection, and certainly before data analysis, real world considerations and interdisciplinary consultations represent a crucial step of building a regression model.

Criterion for a Linear Regression Most of our diagnostic techniques involve specific statistical computations and comparisons, but for some diagnostics, the human brain is an important tool because of a unique ("wet") circuitry that allows for more efficient pattern recognition. The **Criterion for a Linear Regression** is one such technique.

Statistical software is useful for applying this technique, because the first step is to construct a scatterplot of the bivariate data that **does not** include the least squares regression line. For relatively small data sets this can be done by hand, but it is typically much faster computationally.

Once the bivariate data is graphed over the range of the sample, the statistician looks for any strong visual evidence of non-linear relationships and/or any visual evidence that the variance of the response variable from the underlying linear pattern changes with respect to the predictor variable.

When applying the criterion for a linear regression, it is important to understand that a nonlinear trend or tendency for the variability to change should be quite pronounced for one to reject a linear model, especially when small data sets are involved (sometimes are brains push the pattern recognition a little **too** far; this is what I call "seeing constellations").

Residual Error Normality Plots According to model assumptions, the residual errors should have a normal distribution. Assessing normality is something covered in prerequisite courses. For large data sets, one can create a histogram of the residual errors to verify that it has a bell-shaped distribution. For smaller data sets without resolved histograms, normal probability plots are typically used to assess normality. This is done computationally in practice.

Residual Error Plotted Against Predictor Residual errors are assumed to be independent random variables. By plotting the bivariate set of predictor and residual error values, (x_i, ϵ_i), we can assess this independence assumption by looking for any patterns that indicate a correlation of residual errors, which violates model assumptions.

These plots are also used to assess nonlinearity of the regression functions. If residual error is scattered but appears to follow a trend that is not centered about the predictor axis, this can be a sign of a nonlinear relationship.

Further Residual Error Analysis Residual errors provide a rich resource for model assumption analysis. In addition to the preceding two items, absolute (or squared) residuals are often plotted against the predictor to assess constancy of error variance. Other techniques included investigating omitted factors using residual analysis and plots of residual errors vs. fitted values, time, or another sequence.

Hypothesis Tests of Model Assumptions The preceding diagnostic techniques listed above all involve some subjectivity. Model assumptions can also be tested using objective, programmable inferential hypothesis testing techniques. These hypothesis techniques are covered in Module 5 and include tests for normally distributed randomness, constancy of variance, and outliers.

Exercises

Complete the exercises below for your Project 2 data collection and analysis. Use complete sentences for all of your exercise solutions.

Exercise 1 (5 points): Chapter 1 of [2] introduces some initial data analysis methods commonly used for numerically and graphically summarizing data sets. There are five exercises at the end of this introductory chapter (page 12), each of which asks the student to "make a numerical and graphical summary of the data, commenting on any features that you find interesting." and "limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data."

Split these five textbook exercises evenly among each other (at most 3 students per problem) and present your analysis in class on 2/3/21.

Exercise 2 (5 points): Collect a bivariate data set the following way: Choose a word processing document that has at least one full page of writing in a relatively small font. Gradually increase the fonts size and collect the following variables: x_j =font size, y_j =words per page (on page 1). Collect these measurements for at least 15 different font sizes and complete each of the following steps:

(1) Should this be classified as a controlled experiment or an observational study? Explain.
(2) Construct a scatterplot of the data **without** the least squares regression line. Does the data appear to satisfy the criterion for a linear regression? Explain
(3) Construct a normal probability plot of the residual errors. Does it appear that the normality assumption for the error term in the model is satisfied?
(4) Plot the residual values associated with the simple linear regression model as a function of the predictor variable. Based on this plot, does this residual plot suggest the presence of a nonlinear response function? Explain.
(5) Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ using statistical software. Write down the formula for a your least squares regression line and plot the regression line on a scatterplot of the sample data.
(6) Calculate the proportion of variability in words per page (on the first page) that is explained by the linear model relating words per page to font size.

(7) Do you think a simple linear regression model is appropriate for the underlying bivariate population in this problem? Explain.

Be prepared to share your work and back up your conclusions on 2/3/21.

Exercise 3 (10 points): Consider the data set "cheddar" from our textbook. Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid, and each sample was tasted and scored by a panel of judges and the average taste score produced [2].

You can access this data in R as follows:

```
install.packages("faraway")
library(faraway)
head(cheddar)
```

The last command shows the first six rows of the data frame and allows the user to see all the variable names in the data frame.

Split all 6 possible combinations of two of the four variables recorded in this data set (at most 3 students per problem). If taste is one of the variables used, make taste the response variable. Evaluate model assumptions by completing each of the following steps:

(1) Construct a scatterplot of the data **without** the least squares regression line. Does the data appear to satisfy the criterion for a linear regression?
(2) Construct a normal probability plot of the residual errors. Does it appear that the normality assumption for the error term in the model is satisfied?
(3) Plot the residual values associated with the simple linear regression model as a function of the predictor variable. Based on this plot, does it appear that constant variance assumption associate with the model is satisfied?
(4) Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ using statistical software. Write down the formula for a your least squares regression line and plot the regression line on a scatterplot of the sample data.
(5) Calculate $\hat{\sigma^2}$ for your data set using statistical commands in R. Verify that it is equal to $\frac{SSE}{n-2}$ using mathematical commands in R.
(6) Calculate the proportion of variability in the response variable that is explained by the linear model relating your predictor variable and response variable.
(7) Interpret the estimated slope of the regression line in terms of the real world response and predictor variable.
(8) Choose a value of your predictor variable that was not used in your actual sample, call this x^*. Construct the estimated response value, $\hat{y}$ associated with this predictor valuable; that is, estimate $\mu_{Y \cdot x^*}$.
(9) Do you think a simple linear regression model is appropriate for the underlying bivariate population in this problem? Explain.

We will work on these exercise steps during our 2/1/21 and 2/3/21 classes. A report of your final results for this problem can be completed after class.

References

[1] Dunbar, S. *Mathematical Modeling in Economic and Finance: Probability, Stochastic Processes, and Differential Equations*. AMS/MAA Press (2019).

[2] Faraway, J. *Linear Models with R, 2nd ed.*. Chapman and Hall, (2014).
[3] Faraway, J. (April 27, 2018) *Linear Models with R* translated to Python. Blog post. https://julianfaraway.github.io/post/linear-models-with-r-translated-to-python/
[4] Fahrmeir, Kneib, Lang, Marx , *Regression*. Springer-Verlag Berlin Heidelberg (2013).
[5] Kutner, Nachtsheim, Neter, Li *Applied Linear Statistical Models, 5th ed.* McGraw-Hill/Irwin (2005).
[6] Draper, Smith *Applied regression analysis 3rd ed.* Wiley (1998)