Simple Linear Regression

Dr. Austin Brown

Kennesaw State University

Introduction

Some of the materials included in today's lecture have been adapted from those created by Dr Taasoobshirazi (Thank you Dr T!).

Introduction

- Review of the Simple Linear Regression Model
- Estimating the Regression Coefficients (Least-Squares Method)
- Hypothesis Testing of the Regression Coefficients
- Interval Estimation
- Prediction
- Coefficient of Determination

Review of the Simple Linear Regression Model

- Recall from last class, one of the general, overarching goals of regression analysis is to quantify the relationship between two (or more) variables.

- Another way of thinking about: based on our knowledge/experience, we hypothesize that the variability we see in some observable phenomenon of interest (say MLB Team Regular Season Wins) can potentially be explained by some other observable/measurable characteristic (say, total runs scored). It's really a tool to help us better understand why something occurs the way it does.

- For example, suppose we wanted to see how the number of runs a MLB team scores in the regular season explains their total number of wins in that same regular season.

Review of the Simple Linear Regression Model

- In this case, our outcome (or response or dependent variable) is Team Wins and our explanatory variable (or independent variable or regressor or predictor or covariate) is Runs Scored. So our model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

,

- where:
  - y_i is Team i's regular season wins
  - β_0 is the y-intercept (or value of y_i when $x_i = 0$)
  - β_1 is the slope parameter (the expected mean change in y_i for a unit increase in x_i)
  - ε_i is the random error term (the difference between what we predict y_i to be, which is $\beta_0 + \beta_1 x_i$, and what we actually observe y_i to be).

Review of the Simple Linear Regression Model

- I find it handy to sometimes write out the full model with words. This is a useful practice as it can improve clarity when presenting to folks unfamiliar with your study (including in manuscripts).

$$\text{Wins}_i = \beta_0 + \beta_1(\text{Runs})_i + \varepsilon_i$$

- Here, β_0 specifically denotes the number of wins a team would suspect to earn if they scored exactly zero runs (hard to win if you aren't scoring at all haha). β_1 is the total number of wins a team will expect to have for every additional run scored (more meaningful than β_0, contextually).

Review of the Simple Linear Regression Model

- What are our assumptions again?

1. An approximately linear relationship exists between the outcome and explanatory variable.
2. The observations are independent of each other (nature of our sampling method dictates this).
3. The error terms are independently and identically distributed normally with a mean of 0 and a constant variance term:

$$\varepsilon_i \sim N(0, \sigma^2) \forall i$$

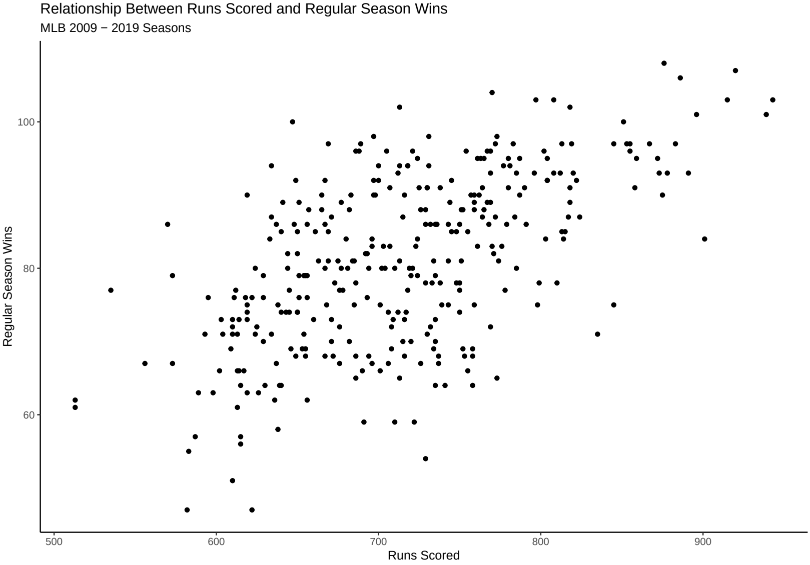
Estimating the Regression Coefficients (Least-Squares Method)

- Okay, so we know the assumptions and we get the overall concept of what a SLR model is trying to accomplish.

- We want to fit a line to our data to help us with understanding the relationship between our two variables as well as for testing, inference, and prediction.

- How do we do that?

Estimating the Regression Coefficients (Least-Squares Method)



Estimating the Regression Coefficients (Least-Squares Method)

- There are lots of different approaches that one could use to draw a straight line through the points. What would make one line better than another?

- It is logical to assume that we would want the line to be as close to the points on the scatterplot as possible, right? Another way of thinking about this is that we would want to draw the line such that the difference between our observed point, y_i, and our predicted point, $\hat{y}_i$ are minimized for all of the i's.

Estimating the Regression Coefficients (Least-Squares Method)

- As is common, instead of minimizing the distance between y_i and $\hat{y}_i$, we instead minimize the square distance:

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- We know that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, so using partial derivatives (and technically partial second derivatives as well), we can find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimize the above function.

Estimating the Regression Coefficients (Least-Squares Method)

- If you were to work it all out, we arrive at what are referred to as the **ordinary least-squares estimators**:

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

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

Estimating the Regression Coefficients (Least-Squares Method)

- So once we go through and actually obtain estimates for our regression coefficients and predicted values, we have what I call the fitted model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Notice that there isn't a residual term here. This is because the fitted model is a line, our estimated mean for each of our y_i's.
- Let's work through an example using the baseball data with R. We'll learn how to do this using the traditional method and then also with the `tidymodels` package.

Estimating the Regression Coefficients (Least-Squares Method)

- Some useful properties of least-squares estimation:

1. $\sum(y_i - \hat{y}_i) = 0$
2. $\sum y_i = \sum \hat{y}_i$
3. The least-squares regression line will always pass through the centroid of the data ($\bar{x}$,$\bar{y}$).
4. $\sum x_i \varepsilon_i = 0$ and $\sum \hat{y}_i \varepsilon_i = 0$.

Hypothesis Testing of the Regression Coefficients

- So we have gone through and estimated our coefficients and built our fitted model. How do we know if the model is significant (i.e., how do we test our regression coefficients)?

- As is the case with all parametric types of hypothesis tests, we have to know what the standard error is for our sample statistic (in this case our sample statistics are $\hat{\beta}_0$ and $\hat{\beta}_1$).

- To do this, we first have to know how to estimate the variance that's supposed to be constant for the residuals. This will aid in our search for the regression coefficients' standard error term.

Hypothesis Testing of the Regression Coefficients

- If part of the goal of SLR is to figure out how well our predictor is explaining the variability in our response, then we need to know how to quantify the variability. That is, how do we estimate σ^2?

- We have a few different metrics we need to know:

$$SSTotal = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSRegression = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSError = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Of note, $SSTotal = SSRegression + SSError$.

Hypothesis Testing of the Regression Coefficients

- Because *SSError* (or *SSE* for shorthand) is quantifying the total squared deviance between our observations and predicted values, it is a natural candidate for estimating σ^2.

- However, *SSE* on its own is sort of like the numerator of s^2. So in order to get an estimate of σ^2 (which I denote $\hat{\sigma}^2$), we divide *SSE* by its degrees of freedom, which for regression is the number of observations less the two regression coefficients we're estimating. This quantity is referred to as "Mean Squared Error (MSE)."

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2}$$

Hypothesis Testing of the Regression Coefficients

- Now that we have an overall estimate of the variance, we can move ahead with hypothesis testing.

- We can test both of the regression coefficients individually using t-tests.

- First, let's test the slope parameter, $\hat{\beta}_1$.

Hypothesis Testing of the Regression Coefficients

- Here, our null and alternative hypotheses are:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- Our test statistic is:

$$t_{stat} = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t(n-2)$$

- where:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- If $|t_{stat}| > |t_{\alpha/2,n-2}|$, then we reject H_0.

Hypothesis Testing of the Regression Coefficients

- The test of $\hat{\beta}_1$ is often called the test of the significance of the regression. Why?

- If $\hat{\beta}_1$ is not significant, then this implies that no strong evidence exists in our data for us to confidently say a relationship likely exists between these two variables. And vice versa.

- Think about the construction of $\hat{\beta}_1$ and how this conclusion makes since ($\hat{\beta}_1$ is a function of the correlation coefficient, r!).

Hypothesis Testing of the Regression Coefficients

- For the y-intercept, $\hat{\beta}_0$, we're testing a very similar set of hypotheses, but we have a slightly different test statistic.

$$H_0 : \beta_0 = 0$$

$$H_1 : \beta_0 \neq 0$$

- The test statistic is:

$$t_{stat} = \frac{\hat{\beta}_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t(n-2)$$

- Like with the $\hat{\beta}_1$ test, we compare this test statistic to a critical value (or find p-value associated with the test statistic) to make a determination about H_0.

Hypothesis Testing of the Regression Coefficients

- Let's go through the baseball example to get our parameter estimates and test statistics using both the traditional and tidy techniques.

Hypothesis Testing of the Regression Coefficients

- Alternatively to going through the t-test method, we can also test the significance of the regression (e.g., $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$) using an omnibus or overall test. This is typically done as a first step in ANOVA models or sometimes multiple linear regression models, too.
- In most statistical softwares, we get an overall ANOVA table as part of the output that has the following form:

Hypothesis Testing of the Regression Coefficients

Table 1: ANOVA Table for Regression

Source	Sums of Squares	DF	Mean Square	F	P
Model	SSR	1	$SSR/1$	MSR/MSE	p
Error	SSE	$n-2$	$SSE/(n-2)$		
Total	SST	$n-1$			

Hypothesis Testing of the Regression Coefficients

- Interestingly, for SLR, the t-test for β_1 and the omnibus F-test will yield identical results because, for SLR the omnibus F-statistic is β_1's t-statistic, squared:

$$t^2_{Stat} = F_{Stat}$$

- In general for you math stats fans, if you square any t distribution (centralized or non-centralized), you'll obtain an F distribution (centralized or uncentralized).

- Let's see this in play with our two running examples.

Interval Estimation

- As mentioned in last class, hypothesis tests and confidence intervals are interrelated and for the same value of α will give you identical conclusions with respect to the null and alternative hypotheses.

- Because we have sampling distributions for our regression coefficients, we can build confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$. For the latter:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{\frac{MSE}{S_{xx}}}$$

Interval Estimation

- For the former:

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}$$

- For the traditional method of model fitting in R, you can use the `confint` function to obtain interval estimates. For the tidymodels method, we can change an argument in the `broom::tidy` function.

- One thing to be mindful of with regards to the interpretation of the confidence intervals: when we say, "we're 95% confident that the true value of the parameter is between the lower and upper limits," that level of confidence has to be considered in the frequentist interpretation of probability.

Interval Estimation

- With frequentist probability, we think about things over the long run. So with a 95% confidence interval, the 95% means that if the state of the world we've observed is the true state (meaning a world in which the null is or isn't false), then if we took a large number of samples and fit a large number of regression models, we'd expect about 95% of those confidence intervals we've constructed to contain the true value of the regression coefficient.

- Our confidence is with respect to the method, not the exact upper and lower boundaries.

Interval Estimation

- This is why some people prefer Bayesian statistics as they interpret probability as a “measuring stick of uncertainty” which is more intuitive despite the methodologies being somewhat less intuitive (at least in my opinion!).

Interval Estimation

- We can also obtain interval estimates for each of our fitted values, $\hat{y}_i$. Since our fitted line is literally an estimate of the mean of y_i, we can think of an interval estimate for our fitted $\hat{y}_i$ values as a confidence interval for the mean. Given an observation for our explanatory variable, say x_i:

$$E[y_i|x_i] = \hat{y}_i \pm t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)}$$

Prediction

- Now, another primary goal associated with regression methods is prediction. Prediction is somewhat different than the process we go through to fit our model because, implicit in prediction is the collection of new data to make a new prediction.

- So we obtain some new observation for our explanatory variable that I'll call, x_0. Our prediction, which I'll denote as $\hat{y}_0^*$, is:

$$\hat{y}_0^* = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Prediction

- and has an associated *prediction* interval:

$$\hat{y}_0^* \pm t_{\alpha/2,n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

Prediction

- Notice, because of the additional “1” term in the standard error for a predicted value, its interval will be wider than that of the fitted values.

- This is because with prediction, we’re predicting a single value with new data and there’s an increase in uncertainty associated with that.

- Don’t forget not to extrapolate!

Coefficient of Determination

- The last thing I want to touch on with SLR is the concept of the Coefficient of Determination (sometimes called, R^2).
- Obviously, if we have a significant regression, then that likely implies that there is likely some sort of linear relationship between our predictor and response.
- However, because our test statistics are functions of the sample size, a large enough sample will give a significant result when a practical one may not be present.

Coefficient of Determination

- In some instances, this may not be so evident. So it then becomes necessary to have alternative ways of assessing the fit of the model. One of the most widely used methods is R^2.

- Mathematically, $R^2 = SSR/SST$. SSR is literally the amount of overall variability "explained" or accounted for by the model.

- Since $SST = SSR + SSE$, the ratio of SSR to SST is then literally the proportion of explained variability. As this number approaches 1, the better our model is doing.

Coefficient of Determination

- One cool thing about R^2 for SLR models is that it is the correlation coefficient squared (so literally, r^2).

- However, we have to be a little bit careful about R^2. When we start adding explanatory variables to our model, the unadjusted R^2 will increase regardless of the relevance of the explanatory variables (so like, the amount of money I've spent at Dunkin likely has no relevance to the number of Wins the Braves got in 2018).

- Further, a good R^2 isn't necessarily an indicator of the appropriateness of a linear model. We have to take a 360 view of everything going on with our model to have a good understanding of model fit/adequacy and can't rely on individual metrics/statistics to tell the whole story.