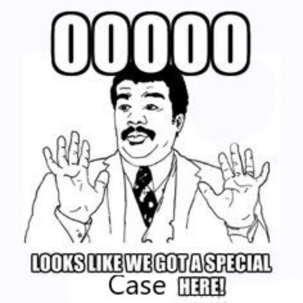
Simple Linear Regression

Javier E. Flores

April 24, 2019

Introduction

At the end of our last lecture, I mentioned that the ANOVA model is simply a special case of regression model.



In this lecture, we will explore the simple regression model (SLR), which (as we know) allows for testing associations between two quantitative variables.

Simple Linear Regression

Given that ANOVA is a special case of regression model, much of the same concepts discussed in that lecture carry over to our discussion of simple linear regression.

Like the ANOVA model, the SLR model involves using a single explanatory variable in order to predict a quantitative outcome variable.

In the SLR model, however, this single explanatory variable is quantitative rather than categorical.

With this change in explanatory variable type (i.e. categorical in ANOVA to quantitative for regression), how do the concepts introduced in our ANOVA discussion change, if at all?

The Null Model and Residuals

In our ANOVA discussion, we learned about the null model, which makes the same prediction for every observation in a dataset.

For the ANOVA model, this prediction is the overall mean regardless of the observation's group membership.

Similarly, in the SLR model, the prediction is the overall mean regardless of the observation's explanatory variable.

> Thinking way back to our LSD and math scores example, this would be the equivalent to saying that regardless of how high your LSD concentration is, you'll do just as well on a math test as anyone else.

In either the ANOVA or SLR case, the key to evaluating each model (null or otherwise) is in the residuals:

$$r_i = y_i - \hat{y}_i$$

Sum of Squares Error

For a given model, the residuals tell us how far off each of our predictions is from the actual observation. In other words, each residual quantifies the amount of error made with each prediction.

For the SLR null model, each prediction, $\hat{y}_i$, is the same and equal to the overall mean, $\bar{y}$. This is also known as an **intercept only model**.

This model may be expressed as $\hat{y}_i = a$, where $a = \bar{y}$. From this expression we see that the model has only one parameter, the intercept a, which means that $d_0 = 1$.

The "alternative" SLR model is $\hat{y}_i = a + bx_i$. This model adjusts the prediction for each observation according to the value of X. This alternative model has two parameters, the slope and intercept (i.e. $d_1 = 2$).

Computing and summing the residuals for each of these model formulations, we obtain the SSEs:

Model	Expression	SSE
Null	$\hat{y}_i = a$	$\sum_i (y_i - a)^2$
Alternative	$\hat{y}_i = a + bx_i$	$\sum_i (y_i - (a + bx_i))^2$

SLR ANOVA table

In the same way that we defined *SSG* in ANOVA, we may define the sum of squares of the alternative model, *SSM*:

$$SSM = SST - SSE$$

With *SSM* and *SST*, we can also compute the coefficient of determination for the SLR model:

$$R^2 = SSM/SST$$

Additionally, we can construct the same ANOVA table we've seen before and use an F-test to evaluate the linear association between X and Y:

Source	DF	SS	MS	F-Value	P-Value
"Model"	$d_1 - d_0$	SSM	MSM	MSM/MSE	Use $F_{d_1 - d_0, n-d_1}$
Error	$n - d_1$	SSE	MSE		
Total	$n - d_0$	SST			

Example

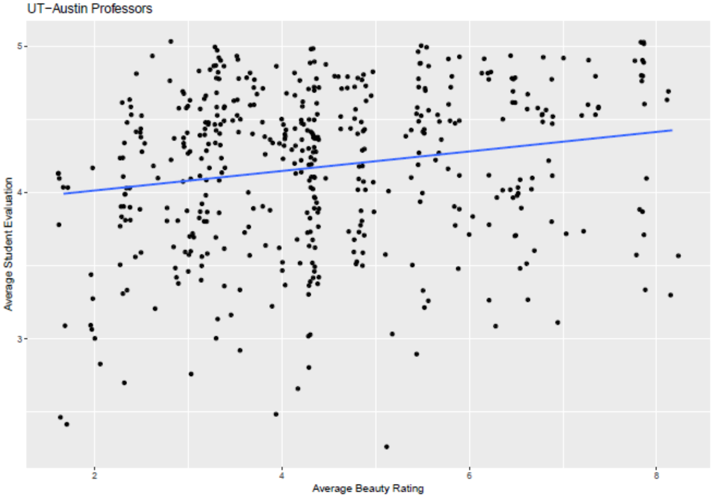
For every university course, students are asked to complete evaluations which ask a variety of questions related to the course content and how well the instructor did in teaching it.

At the University of Texas at Austin, a study was interested in determining whether there was a relationship between end of course evaluations and the "beauty rating" of a professor.

Six students (three male and three female) were asked to give each UT-Austin professor a beauty rating on a 1-10 scale based on a provided photograph of each.

Using these data, a model was constructed relating the average student evaluation score (which was on a 5 point scale) to the average beauty score.

Example


UT−Austin Professors
Average Student Evaluation
5
4
3
Average Beauty Rating
2
4
6
8

Practice

With your groups, load the UT Profs dataset (available on the course website) into Minitab and answer the following questions:

1) What value would the null model predict for each professor's average evaluation score (i.e. "score")?
2) Fit a simple linear regression model that uses "avg_beauty" to predict "score". Using the resulting ANOVA table, does knowing a professor's average beauty rating significantly improve how well you can predict their average evaluation score?
3) Using the previous model, predict the evaluation score for a professor with an average beauty rating of 4.418. Have you seen the resulting prediction before?

Solution

1) The null model using the sample mean for every prediction. Therefore the null model prediction would be 4.17, which is the sample mean of the average evaluation scores for professors.

2) The F-statistic is 16.7. Using an F distribution with degrees of freedom of 1 and 461, we obtain a p-value of 0.00005 which indicates that beauty ratings are predictive of evaluation scores. Since the slope coefficient is positive, we conclude that higher beauty ratings are associated with higher evaluation scores.

3) The predicted value for a beauty rating of 4.418 is 4.17, which is the same as the null model prediction for all professors.

Inference for Slope

As we did with ANOVA, we should perform follow-up analyses after a statistically significant F test.

In the case of the ANOVA model, these follow-up analyses involved performing pairwise comparisons across group means in order to find where difference(s) arose specifically.

In the SLR model, follow-up analyses of interest involve the regression slope:

$$\text{Confidence interval for } \beta \quad b \pm t_{crit} SE$$

$$\text{Test of } H_0 : \beta = 0 \quad t_{test} = \frac{b - 0}{SE}$$

For each of these inferential procedures, we use a t-distribution with $n - 2$ degrees of freedom.

More Rigor!

Up until this point, I've been expressing the SLR model as:

$$\hat{y}_i = a + bx_i$$

While this expression communicates the practical intuition behind regression, it does not properly characterize the true regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

What is the difference between these two models?

The first, "intuitive" model describes the fitted regression model. As such, *a* and *b* are *estimates* of the true population parameters (α and β) that are used in the expression for the true regression model.

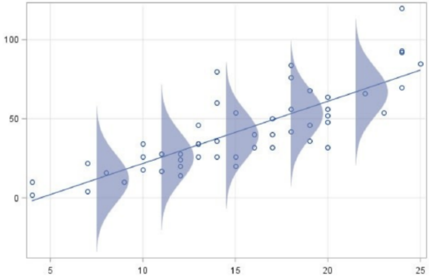
Additionally, the first model only provides the resulting prediction, $\hat{y}_i$, and not how it deviates from the observed data point. In contrast, the more rigorous formulation captures the deviation, or error, through ϵ_i.

More on ϵ

In the formal regression model, we assume that the ϵ_i (i.e. errors) are independent and are all identically normally distributed.

The ϵ_i represent the population errors and are estimated by our model residuals, r_i.

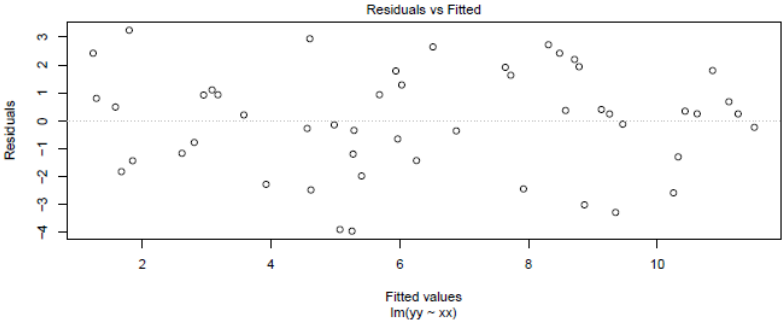
Therefore, we can use our model residuals to determine whether the underlying assumptions of our model (i.e. independent and identically normal errors) are met.



Residual vs. Fitted Value Plots

A common way to assess model assumptions is to plot the model residuals,r_i, against the fitted (i.e. predicted) values, $\hat{y}_i$ of a model.

If errors are identically distributed (one of the key SLR assumptions), you should see a plot like the one below in which the residuals occur in either direction, in similar magnitude, regardless of the predicted value.

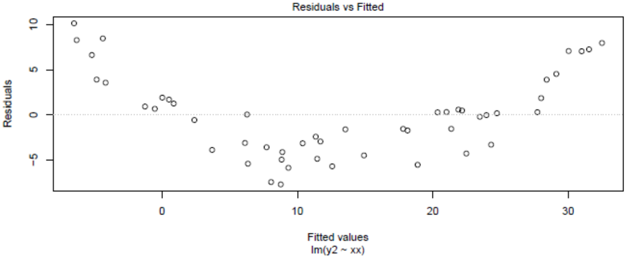


Residual vs. Fitted Value Plots

In contrast to the previous plot in which the errors were randomly scattered about 0, patterns in the residual plot are indicative of systematic error.

In other words, a pattern in the residual indicates that your model is poor and does not fit well.

In the plot below, we see that our regression model ($\hat{y} = a + bx$) will over-predict at large or small values of x and under-predict at intermediate values.

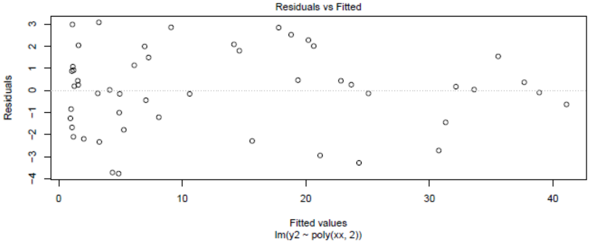


Residual vs. Fitted Value Plots

To remedy the lack of fit seen previously, we might consider adding a quadratic component to our model (as suggested by the U-shape in the fitted v. residual plot).

Doing so results in the *multiple regression* model:
$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$

This model is a multiple regression model because we are using two quantitative predictors, X and X^2, as opposed to one. The number of parameters in this model is 3.



Practice

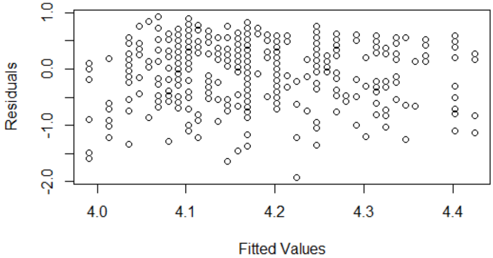
With your groups,

1) Fit a simple linear regression model using "avg_beauty" to predict "score" with the UT Profs dataset.
2) Assess the fit of this model using a residual vs. fitted value plot. Are the residuals identically distributed?
3) Fit a quadratic model, compare this model's coefficient of determination with the simple linear model.
4) Fit a cubic model, compare this model's coefficient of determination with the other two models, and decide which is best.

Solution

Looking at the residual vs. fitted value plot, the residuals seem to be randomly dispersed, which is good.

The simple linear regression model has an R^2 of 3.5%, the quadratic has an R^2 of 4.39%, and the cubic has an R^2 of 6.1%. It appears the cubic model may be best, but it could be overfit.



Comparing Nested Models

In the previous example, we were limited to comparing models based on their R^2.

Given that R^2 always increases with added complexity, this is not a reliable way to decide which model is best from among a group of models.

Instead, we can generalize concepts in ANOVA in order to make comparisons between any two **nested models**.

Consider the following models from the previous example:

$$\begin{aligned} M_1 &: Y = \alpha + \beta_1 X \\ M_2 &: Y = \alpha + \beta_1 X + \beta_2 X^2 \\ M_3 &: Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \end{aligned}$$

M_1 is a special case of M_2 (i.e. take $\beta_2 = 0$), and M_2 is a special case of M_3 (i.e. take $\beta_3 = 0$). Therefore, M_1 is nested in M_2 which is nested in M_3.

Question: Is the null model nested in M_1?

Comparing Nested Models

The null model ($M_0 : Y = \alpha$) is indeed nested in M_1!

This considered, we compare the fits of any two nested models the same way we would compare the fits of the null (M_0) and alternative (M_1) models in ANOVA (i.e. using an F test).

The only difference is that d_0 and d_1, the number of parameters in our "null" and "alternative" models respectively, changes depending on how we define the "null" model. Comparing M_1 and M_2, for example, we would say $d_0 = 2$ since M_1 has two parameters and $d_1 = 3$ since M_2 has three parameters.

Source	DF	SS	MS	F-Value	P-Value
"Model"	$d_1 - d_0$	SSM	MSM	MSM/MSE	Use $F_{d_1 - d_0, n-d_1}$
Error	$n - d_1$	SSE	MSE		
Total	$n - d_0$	SST			

Practice

With your groups,

1) Fit a simple linear regression model using ”avg_beauty” to predict ”score” with the UT Profs dataset and record the SSE.

2) Fit a quadratic model and construct the ANOVA table necessary to conduct an F-test to determine if the added complexity of a quadratic term significantly improves the model.

Hint: The SSE of the ”null” model should be used as the SST in the ANOVA table in 2). Similarly the SSE degrees of freedom for the ”null” model should be used as the SST degrees of freedom in the ANOVA table in 2).

Solution

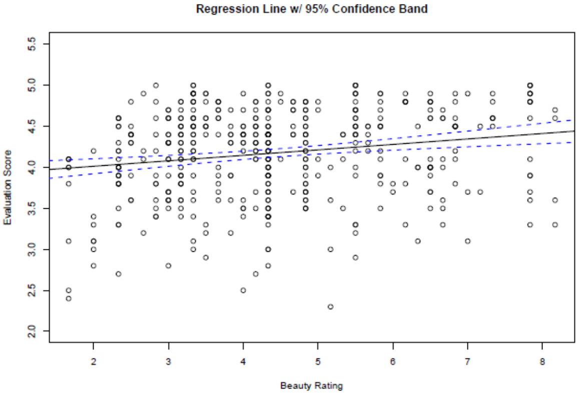
The resulting ANOVA table, which compares the simple regression model to the one which includes a quadratic term, is shown below:

Source	DF	SS	MS	F-Value	P-Value
Model	1	1.2	1.2	4.21	0.041
Error	460	130.7	0.285		
Total	461	131.9			

The reason we use the SSE from our non-quadratic fit in the table here is because our interest is in determining whether adding a quadratic term helps explain any of the remaining unexplained variability (i.e. SSE) in our outcome after modeling with a single linear predictor. In this case, adding a quadratic term does indeed improve the model fit (since $0.041 < 0.05$).

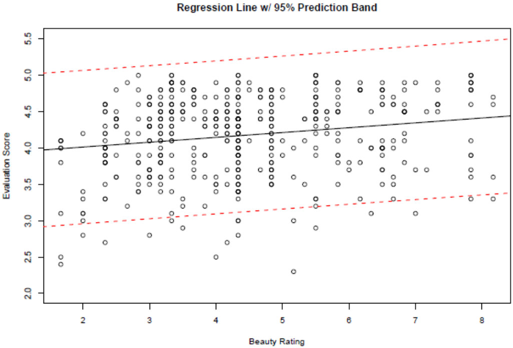
Confidence and Prediction Intervals

Fitted regression models are typically accompanied by a confidence interval, prediction interval, or both. Shown below is a regression plot with the confidence band (i.e. interval) plotted with hyphenated blue lines.



Confidence and Prediction Intervals

Shown below is a regression plot with the prediction band (i.e. interval) plotted with hyphenated red lines.



Question: What makes prediction and confidence intervals different?

Confidence and Prediction Intervals

Consider a specific value of the explanatory variable, i.e. $X = x^*$ where x^* denotes the specific value of the explanatory variable X.

The confidence interval describes the chances of capturing the mean response for all members of the population with $X = x^*$.

The prediction interval describes the chances of capturing the individual response for a given individual with $X = x^*$.

For example, the 95% confidence interval indicates that we are 95% confident that the mean evaluation score for a professor with a beauty rating of 7.8 is between 4.1 and 4.4.

On the other hand, the 95% prediction interval indicates we are 95% confident that an individual professor with a beauty rating of 7.8 has an evaluation score between 3.2 and 5.4.

Practice

With your groups,

1) Fit a simple linear regression model using "avg_beauty" to predict "score" with the UT Profs dataset.

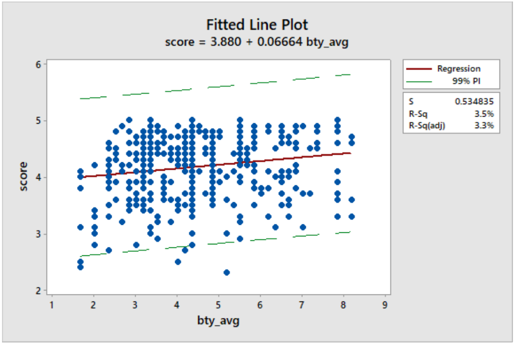
2) Use Minitab to plot the regression model with a 99% prediction interval.

3) What proportion of the observed data do you expect to be within the prediction band? What proportion is actually within the prediction band?

Solution

Since we are constructing a 99% prediction interval, we would expect 99% of our observed data to be contained within the prediction band.

The figure below shows that about 5/463 datapoints were excluded by the prediction bands, suggesting that 98.9% of the observed data were actually in the prediction band.



Summary

Simple linear regression is a statistical model that uses a quantitative explanatory variable to predict a quantitative outcome.

As in ANOVA, we can assess the propriety of this model (i.e. how well it fits) using an F-test.

Following an F-test, it is also important to conduct inference on the model parameters themselves, namely the slope parameter β.

We can also use confidence and prediction intervals to conduct statistical inference on the predictions made by a model.

Summary

We can also generalize ANOVA concepts to compare nested models.

This technique is particularly important in the context of multiple regression, in which regression models have several predictors.

In the last lecture of this semester, we will further explore multiple regression and increase our understanding of comparing several models.

Wrap-Up

Right now, you should...

- Understand the similarities and differences between regression and ANOVA
- Understand how to evaluate a regression model using an F-test
- Make inferences about the regression slope parameter using a t-test or confidence interval
- Know the difference between confidence and prediction intervals

These notes cover chapter 9 the textbook. Please read through the section and its examples along with any links provided in this lecture.