Inference with Simple Regression

Alan B. Gelder
06E:071, The University of Iowa[1]

1 Introduction

Moving to infinite means: In this course we have seen one-mean problems, two-mean problems, and problems with several means (ANOVA). Regression falls right into this pattern by comparing an infinite number of means. **(How?)**

Where are the bell curves? From the analysis of *variance* we learned that means are only one part of the picture. We also need to have a good understanding of the underlying distribution. Pages 571 and 572 of the text provide good illustrations of the bell curves in regression.

Population and sample regression lines: In single mean problems, our goal is to determine what the true population mean μ is based on our sample. In regression, we want to identify the true regression line between the x and y variables. Since a line is fully characterized by its intercept and slope, our goal is then to identify the true intercept β_0 and true slope β_1 of the regression line. The population (or true) regression line for *an individual* is written as follows:

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

The ε_i is an error term which we assume is normally distributed. This is where the bell curves for regression come in. For a given value of x there is a distribution of points that y can come from, which is centered on the regression line. Although individual points may not be exactly on the regression line, the regression line will be correct on average. The regression line for the *average* value of y for a given x is shown below:

$$\mu_y = \beta_0 + \beta_1 x$$

Based on our sample, we can estimate the true population intercept and the slope of the regression line. The sample regression equation is

$$\hat{y}_i = b_0 + b_1 x_i$$

Inference: In single mean problems, we did confidence intervals and hypothesis testing to infer from the sample information about the true population mean μ.

[1]The notes for this class follow the general format of Blake Whitten's lecture notes for the same course.

In regression we want to similarly be able to infer characteristics about the true intercept β_0 and the true slope β_1. We also want to be able to make statements about where the actual value of y may be for a given x.

2 Dow Jones Example

Table 11.1 in the text lists the assets, sales, and profits in 1999 (given in billions of dollars) of the 30 companies which made up the Dow Jones Industrial Average at the close of that year. With this data we can analyze how a company's assets and sales are related to profits.

2.1 Predicting Profits from Assets

The regression output for predicting profits with assets is given below:

```
Profits = 3.27 + 0.0109 Assets


Predictor      Coef   SE Coef     T      P
Constant     3.2664    0.5763  5.67  0.000
Assets     0.010875  0.003267  3.33  0.002


S = 2.61576   R-Sq = 28.4%   R-Sq(adj) = 25.8%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       1   75.833  75.833  11.08  0.002
Residual Error  28  191.581   6.842
Total           29  267.414
```

This regression equation indicates that for every additional billion dollars of assets a firm holds, its profits increase by 10.875 million dollars on average. That is the interpretation of the slope coefficient b_1. However, we want to know how trustworthy b_1 really is. This will require a formal hypothesis test.

Are assets linearly related to profits? Test using $\alpha = 0.05$.

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

$$t = \frac{b_i}{SE_{b_i}} = \frac{0.010875}{0.003267} = 3.33 \qquad \text{p-value} = 0.002$$

Degrees of Freedom = n − p − 1 = 30 − 1 − 1 = 28. p is the number of regressors (x variables). In simple regression, p = 1.

Reject H_0 since p-value = $0.002 < 0.05 = \alpha$.

There is sufficient evidence to show that assets are linearly related to profits.

We could also test to see if assets are positively related to profits (H_A: $\beta_1 > 0$) or negatively related to profits (H_A: $\beta_1 < 0$). However, it is important to remember that the p-value that Minitab and other statistical software provide in the regression output is the two-tailed test. **Minitab always tests H_A: $\beta_1 \neq 0$.**

What is the t-statistic if H_A: $\beta_1 > 0$?

What is the p-value if H_A: $\beta_1 > 0$?

2.2 Predicting Profits with Sales

We will now turn to the relationship between sales and profits (in billions of dollars). Are sales *positively* related to profits? Test using $\alpha = 0.01$.

H_A: $\beta_1 > 0$
H_0: $\beta_1 \leq 0$

$$t = \frac{b_i}{SE_{b_i}} = \frac{0.03763}{0.01113} = 3.38 \qquad \text{p-value} = 0.001$$

Reject H_0 since p-value = $0.001 < 0.01 = \alpha$.

There is sufficient evidence that sales are positively related to profits.

Hypothesis tests answer a yes/no question, while confidence intervals tell us how much. In addition to knowing that sales are positively related to assets, we want

```
Profits = 2.51 + 0.0376 Sales


Predictor      Coef  SE Coef     T      P
Constant     2.5114   0.7203  3.49  0.002
Sales       0.03763  0.01113  3.38  0.002


S = 2.60431   R-Sq = 29.0%   R-Sq(adj) = 26.4%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       1   77.506  77.506  11.43  0.002
Residual Error  28  189.908   6.782
Total           29  267.414
```

a confidence interval to tell us how much an increase in sales will raise profits on average.

Confidence interval for the b_i:

$$b_i \pm t^* SE_{b_i}$$

Note that this formula can be used for the intercept or slope. In multiple regression, we can use this formula to construct a confidence interval for the slope of any of our predictor variables (as well as for the intercept).

With 95% confidence, how much will profits increase on average if sales increase by one billion dollars?

$$b_1 \pm t^*_{95\%,\, df=28}\, SE_{b_i} = 0.03763 \pm 2.048(0.01113) = (0.01484,\ 0.06042)$$

Interpretation: With 95% confidence, for every one billion dollar increase in sales, profits increase between 14.8 and 60.4 million dollars, on average.

2.3 CI and PI for Sales

The examples so far have focused on doing hypothesis tests and confidence intervals for the slope of the regression equation β_1 (the hypothesis tests and confidence

intervals for the intercept β_0 are similar). However, in addition to the slope and the intercept, will also want to have some tools for evaluating the prediction of our regression equation $\hat{y}$.

Confidence intervals for $\hat{y}$: The regression equation predicts a value for y based on a value of x. The idea of a confidence interval for $\hat{y}$ is that we take all of the possible x variables that have a particular value, x^*, and then we calculate a confidence interval for $\hat{y}$ based on this entire group of x variables at x^*.

For instance, think of all of the firms in the world that make 65 billion dollars of sales in a year. What is a 95% confidence interval for profits for that entire group of firms?

The following commands in Minitab can be used to find the answer to that question:

Regression > Regression > Response: Profits; Predictors: Sales > Options > Prediction intervals for new observations: 65; Confidence level: 95 > OK > OK

This is the bottom of the regression output:

```
Predicted Values for New Observations

New Obs    Fit  SE Fit      95% CI           95% PI
      1  4.957   0.509  (3.914, 6.001)  (-0.478, 10.393)


Values of Predictors for New Observations

New Obs  Sales
      1   65.0
```

The formula for a confidence interval for $\hat{y}$ is

$$\hat{y} \pm t^* SE_{\hat{\mu}}$$

Fit $= \hat{y} = 4.957$

SE Fit $= SE_{\hat{\mu}} = 0.509$

$t^* = t^*_{95\%,\, df=28} = 2.048$

95% CI $= \hat{y} \pm t^* SE_{\hat{\mu}} = (3.914,\ 6.001)$

Interpretation: With 95% confidence, the *average* annual profits of *all* firms that make 65 billion dollars in annual sales is between 3.914 and 6.001 billion dollars.

The full formula for $SE_{\hat{\mu}}$ is

$$SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Here, s is the regression standard error and x^* is the point that $\hat{y}$ is being evaluated at.

Prediction intervals for $\hat{y}$: Confidence intervals for $\hat{y}$ are based on an entire *population* of the x variable that have the value x^*. A prediction interval, on the other hand, only relies on a *single object* that has the value x^*.

Predict the profits of a specific firm that makes 65 billion dollars of sales a year with 95% confidence.

The Minitab output shows the answer to this question: $(-0.478, 10.393)$.

Interpretation: With 95% confidence, the profits of a *particular* firm that makes 65 billion dollars in annual sales is between negative 478 million dollars and positive 10.393 billion dollars.

The formula for a prediction interval is

$$\hat{y} \pm t^* SE_{\hat{y}}$$

where

$$SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

There is quite a bit of similarity between $SE_{\hat{y}}$ and $SE_{\hat{\mu}}$. In fact, the only difference between the two standard errors is that $SE_{\hat{y}}$ has an extra 1 in the square root.

Questions: Which is larger: $SE_{\hat{y}}$ or $SE_{\hat{\mu}}$? Which is more precise for a given x^*: a prediction interval or a confidence interval? Does this make intuitive sense?

Unfortunately, Minitab does not display the value for $SE_{\hat{y}}$ in the output.

3 Weight/Pulse Example

Can a person's pulse be predicted based on their weight? Test using $\alpha = 0.05$. Pulse is measured in beats per minute and weight is given in kilograms.

```
Pulse = 44.0 + 0.417 Weight


Predictor     Coef  SE Coef     T      P
Constant     43.96    13.28  3.31  0.002
Weight      0.4172   0.2090  2.00  0.053


S = 9.14600   R-Sq = 9.7%   R-Sq(adj) = 7.3%


Analysis of Variance

Source           DF       SS      MS     F      P
Regression        1   333.28  333.28  3.98  0.053
Residual Error   37  3095.02   83.65
Total            38  3428.31
```

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

$t = 2.00$, and p-value $= 0.053$

Fail to reject H_0 since p-value $= 0.053 > 0.05 = \alpha$.

There is insufficient evidence to show that pulse and weight are linearly related to each other.

Predict the pulse rate for a person who weighs 55 kilograms with 90% confidence.

Predict the pulse rate for all people who weigh 55 kilograms with 90% confidence.

Can we trust these predictions?

Question: Based on the regression output, how can we determine the sample size?

4 Regression Assumptions

In order for least-squares regression to work properly, all of the following assumptions must be met.

A.1 There is a linear relationship between x and y. That is, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

A.2 The error terms are independent of x. Or rather, $E[\varepsilon_i | x_i] = 0$.

A.3 The error terms are not correlated with themselves (i.e. no autocorrelation). Also, the error terms have the same variance σ^2 (i.e. homoscedastic errors).

A.4 The error terms are distributed normally (i.e. $\varepsilon_i \sim N(0,\, \sigma^2)$).

When we study multiple regression we will add another assumption.

A.5 The x variables are linearly independent. In other words, no x variable can be formed as a linear combination of the other x variables.

As a technical note, this last assumption does play a role in simple regression since the intercept is regarded as another x variable.

Notice that three of the five assumptions address properties of the error terms ε_i. Assumption **A.2** is commonly violated with panel data.[2] In time series data, the error terms are typically correlated with themselves, which violates the first part of Assumption **A.3**.[3] In various contexts, the error terms are not distributed normally, and Assumption **A.4** is violated.

There are a variety of techniques that can be applied when one or more of the re-

[2]Panel data follows characteristics about several individuals over time. An individual does not need to be a person. It could be a company, a machine, a country, etc.

[3]Time series data follows a single individual over time. Again, an "individual" is not limited to a person. For instance, following a stock price over time is an example of a time series data set.

gression assumptions is violated. We will address a couple of these techniques in this class; others will be saved for more advanced classes.

4.1 Checking residuals

There is a quick check that we can do to see if Assumptions **A.3** is satisfied. If this assumption is satisfied, then the error terms should all have the same variance and not be correlated with each other. Stated differently, if the assumption is satisfied, then the residuals should not have a distinct pattern. We simply need to check the residuals to make sure that there is *no* distinct pattern in the residuals.

What kinds of patterns in the residuals are we trying to avoid?

Draw an example of residuals that are autocorrelated (or self correlated).

Draw an example of residuals that are heteroscedastic (meaning "different randomness" or different amounts of variance).

In Minitab, we can check for patterns in the residuals by using the following commands:

Stat > Regression > Regression > (Click "Response (Y)" or "Predictor (X)" window) > (Click variables/columns in left window for Response (Y) and for Predictors (X)) > Select

Graphs > Individual Plots > (Click "Residuals versus fits") > OK > OK

Remember, the residuals are simply the difference between the actual value for y (from the data) and the predicted value for y (from the regression: $\hat{y}$):

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

4.2 Transforming Data

One of our assumptions for regression is that the relationship between x and y is linear. Can we do anything with regression if we have a strong relationship between two variables that is not linear? The short answer is yes, but depending on the nonlinear relationship between x and y we may need some complex tools (many of

which are beyond the scope of this class). There are, however, some nonlinear patterns that are easy to transform into linear patterns. For instance, it is not unusual to have data that exhibits exponential growth. If we have exponential growth, we can restore linearity by taking the natural logarithm of the y variable data.

Consider the following example (see Exercise 10.31):

Dynamic random access memory chips have grown dramatically in their capacity (as measured in bits). Data from 1971 through 2000 is shown below:

Year	Bits
1971	1024
1980	64000
1987	1024000
1993	16384000
1999	256000000
2000	512000000

Transform the data in order to obtain a regression equation which predicts bits based on year.

Based on the regression equation, how many bits did a memory chip have on average in 1995?

5 Formulas

$$b_1 = r\frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$t_{(df=n-p-1)} = \frac{b_i}{SE_{b_i}}$$

where p = # of regressors (in simple regression, p = 1)

$$b_i \pm t^* SE_{b_i}$$

$$\hat{y} \pm t^* SE_{\hat{\mu}}$$

$$\hat{y} \pm t^* SE_{\hat{y}}$$