STAT5102

Regression in Practice

II. Remarks and Examples – Simple Linear Regression

Department of Statistics
The Chinese University of Hong Kong

II. Remarks and Examples - Simple Linear Regression

In this chapter, we shall cover

- Further remarks and examples for simple linear regression

Properties of Fitted Regression Line

- The sum of residuals is zero.
- The sum of squared residuals is a minimum.
- The sum of the observed values equals the sum of the fitted values.
- The sum of the weighted residuals is zero when the residual in the i^{th} trial is weighted by the level of the predictor variable in the i^{th} trial.
- The sum of the weighted residuals is zero when the residual in the i^{th} trial is weighted by the fitted value of the response variable for the i^{th} trial.
- The regression line always goes through the point $(\bar{X}, \bar{Y})$.

Some Concerns

- *Regression analysis used to make inferences for the future:* validity of the regression application depends upon whether basic causal conditions in the period ahead will be similar to those in existence during the period upon which the regression analysis is based.

- *Predicting new observations on Y:*
  the predictor variable X itself often has to be predicted.

- *Prediction outside the range of observations:*
  how far beyond the range of past data?

Some Concerns

- A statistical test that leads to the conclusion that $\beta_1 \neq 0$ does not establish a cause-and-effect relation between the predictor and response variables.

- Single versus multiple inferences

- X may subject to measurement errors: the resulting parameter estimates are no longer unbiased.

Example 1

Parameter	Estimated Value	95% Confidence Intervals
Intercept	7.43119	(-1.18518, 16.0476)
Slope	0.755048	(0.452886, 1.05721)

The student concluded from these results that there is a linear association between Y and X. Is the conclusion warranted? What is the implied level of significance?

Yes, the conclusion is warranted and the level of significance is 0.05 ($a = 0.05$).

Someone questioned the negative lower confidence limit for the intercept, pointing out that dollar sales cannot be negative even if the population in a district is zero. Discuss.

The inference might not be appropriate when the model is extended to $X = 0$. In fact, the confidence interval does not necessarily provide meaningful information. In this case, a reasonable approximation would be to adjust the interval to (0, 16.0476).

Example 2

The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X).

Example 2

Obtain a 99 percent confidence interval for β_1. Interpret your confidence interval. Does it include zero? Why might the director of admissions be interest in whether the confidence interval includes zero?

Parameter Estimates									
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**	**95% Confidence Limits**	
Intercept	Intercept	**1**	2.11405	0.32089	6.59	<.0001	0	1.27390	2.95420
X	ACT	**1**	0.03883	0.01277	3.04	0.0029	0.26948	0.00539	0.07227

From the SAS output, the 99% confidence interval for β_1 is (0.00539, 0.07227).

Alternative method: $t_{(.005,\ 118)} = 2.61814$, therefore, the confidence interval for β_1 is $(0.03883 - 2.61814(0.01277),\ 0.03883 - 2.61814(0.01277)) = (0.0054, 0.07226)$

The interval does not contain 0, hence the director of admissions has a statistically significant result indicating that there is a positive linear relationship between GPA and ACT.

Example 2

Test whether or not a linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y). Use a level of significance of .01.

Hypotheses:

Null Hypothesis $H_0: \beta_1 = 0$

Alternative Hypothesis $H_1: \beta_1 \neq 0$

Decision Rule:

Reject the null hypothesis if the test statistics $|T| > t_{(.005,\ 118)} = 2.61814$

Test Statistic:

$|T| = (0.03883 - 0)/0.01277 = 3.04072 > 2.61814$

Conclusion:

Reject the null hypothesis. With 0.01 level of significance, we claim that there is a linear association between student's ACT score and GPA at the end of the freshman year.

Example 2

What is the P-value of your test in part (b)? How does it support the conclusion reached in part (b)?

P-value = 0.0029 (from SAS output)

The P-value is less than 0.01 and this is an expected finding since the null hypothesis is tested at a level of significance 0.01. Further, the P-value is a lot less than 0.01, providing us a strong evidence against the null hypothesis.

Example 3

The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, X is the number of copiers service and Y is the total number of minutes spent by the service person. Assume that first-order regression model is appropriate.

Example 3

Estimate the change in the mean service time when the number of copiers serviced increases by one. Use a 90 percent confidence interval. Interpret your confidence interval.

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**90% Confidence Limits**	
Intercept	Intercept	**1**	-0.58016	2.80394	-0.21	0.8371	-5.29378	4.13347
X	# of copiers serviced	**1**	15.03525	0.48309	31.12	<.0001	14.22314	15.84735

From the SAS output, the 90% confidence interval for β_1 is (14.22314, 15.84735).

Alternative method: $t_{(.05,\ .43)} = 1.6811$, therefore, the confidence interval for β_1 is $(15.0352 - 1.6811(0.4831),\ 15.0352 + 1.6811(0.4831)) = (14.22314, 15.84735)$.

Conclusion: When the number of copiers serviced increases by one, the estimated change in the mean service time is (14.22314, 15.84735) with 90 percent confidence level.

Example 3

Conduct a t-test to determine whether or not there is a linear association between X and Y here; control the a risk at .10. State the alternatives, decision rule, and conclusion. What is the P-value of your test?

Hypotheses

Null Hypothesis $H_0: \beta_1 = 0$

Alternative Hypothesis $H_1: \beta_1 \neq 0$

Decision Rule

Reject the null hypothesis if the test statistics

$|T| > t_{(.05, 43)} = 1.6811$

Test Statistics

$|T| = (15.0352 - 0)/0.4831 = 31.122 > 1.6811$

Conclusion

Reject the null hypothesis. With 0.1 level of significance, we claim that there is a linear association between X and Y.

P-value: less than 0.0001

Example 3

Are your results in parts (a) and (b) consistent? Explain.

Yes. Since both the upper and lower limits are positive (zero is not contained in the interval), with the same α level, the null hypothesis should also be rejected.

Example 3

The manufacturer has suggested that the mean required time should not increase by more than 14 minutes for each additional copier that is serviced on a service call. Conduct a test to decide whether this standard is being satisfied by Tri-City. Control the risk of a Type I error at .05. State the alternatives, decision rule, and conclusion. What is the P*-value of the test?*

Hypotheses

Null Hypothesis H_0: $\beta_1 \leq 14$

Alternative Hypothesis H_1: $\beta_1 > 14$

Decision Rule

Reject the null hypothesis if the test statistics $T > t_{(.05, 43)} = 1.6811$

Test Statistics:

$T = (15.0352 - 14)/0.4831 = 2.1428 > 1.6811$

Conclusion

Reject the null hypothesis. With 0.05 level of significance, we believe that the standard is not being satisfied by Tri-City. *P*- value: $P(T > 2.1428) = 0.0189$.

Example 3

Does b_0 give any relevant information here about the "start-up" time on calls – i.e., about the time required before service work is begun on the copiers at a customer location?

No, since the estimate is zero which does not correspond to any sensible measurement of time.

Example 4

(refer to Copier maintenance problem in Example 3)

Obtain a 90 percent confidence interval for the mean service time on calls in which six copiers are serviced. Interpret your confidence interval.

We need to estimate the mean response $E(Y_6)$ by $\hat{Y}_6$ where
$\hat{Y}_6 = b_0 + b_1(6) = 89.6313$

Two-sided $100(1-\alpha)a\%$ C.I. for $E(Y_6)$ is

$$\left(\hat{Y}_6 - t_{\alpha/2,n-2}s\{\hat{Y}_6\}, \quad \hat{Y}_6 + t_{\alpha/2,n-2}s\{\hat{Y}_6\}\right)$$

Now,

$$t_{\alpha/2,n-2} = t_{.05,43} = 1.6811$$

$$s\{\hat{Y}_6\} = \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]} = \sqrt{79.45063\left[\frac{1}{45} + \frac{(6-5.1111)^2}{340.4412}\right]}$$

Example 4

So, two-sided 90% C.I. for $E(Y_6)$ is

$$\left(\hat{Y}_6 - t_{\alpha/2,n-2}s\{\hat{Y}_6\}, \quad \hat{Y}_6 + t_{\alpha/2,n-2}s\{\hat{Y}_6\}\right) = (87.2838, 91.9788)$$

With a 0.90 confidence level, we estimate the mean service time on calls in which six copiers are service to be between 87.2838 minutes to 91.9788 minutes.

Example 4

Obtain a 90 percent prediction interval for the service time on the next call in which six copiers are serviced. Is your prediction interval wider than the corresponding confidence interval in part (a)? Should it be?

Let the service time on the next call in which six copiers are serviced be $Y_{6(new)}$ which is estimated by $\hat{Y}_6$.

Two-sided 100(1-α)% prediction interval for $Y_{6(new)}$ is

$$\left(\hat{Y}_6 - t_{\alpha/2,n-2}s\{pred\},\quad \hat{Y}_6 + t_{\alpha/2,n-2}s\{pred\}\right)$$

Now,

$$t_{\alpha/2,n-2} = t_{.05,43} = 1.6811$$

$$s\{pred\} = \sqrt{MSE\left[1+\frac{1}{n}+\frac{(X_h-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]} = \sqrt{79.45063\left[1+\frac{1}{45}+\frac{(6-5.1111)^2}{340.4412}\right]} = 9.0222$$

Example 4

So, a two-sided 90% prediction interval for $Y_{6(new)}$ is

$$\left(\hat{Y}_6 - t_{\alpha/2,n-2}s\{pred\}, \quad \hat{Y}_6 + t_{\alpha/2,n-2}s\{pred\}\right) = (74.4641, \quad 104.7985)$$

With a .90 confidence level, we predict that the service time on the next call in which six copiers will be serviced to be between 74.4641 minutes to 104.7985 minutes.

The prediction interval is much wider than the confidence interval in part (a). This is reasonable since the variance is larger which accounts for the variation from item to item.

Example 4

Management wishes to estimate the expected service time per copier on calls in which six copiers are serviced. Obtain an appropriate 90 percent confidence intervals by converting the interval obtained in part (a). Interpret the converted confidence interval.

87.2838/6 = 14.5473, 91.9788/6 = 15.3298. Thus, the 90 percent confidence interval for the expected service time per copier on calls in which six copiers are serviced is (14.5473, 15.3298).

Example 5

A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules.

The data, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y).

Assume that first-order regression model is appropriate.

Example 5

Set up the ANOVA table. Which elements are additive?

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	160.00000	160.00000	72.73	<.0001
Error	8	17.60000	2.20000		
Corrected Total	9	177.60000			

The ANOVA table is given above with Sum of Squares and Degrees of Freedom being additive.

Example 5

Conduct an F test to decide whether or not there is a linear association between the number of times a carton is transferred and the number of broken ampules; control the a risk at .05.

Hypotheses

Null Hypothesis $H_0: \beta_1 = 0$

Alternative Hypothesis $H_1: \beta_1 \neq 0$

Decision Rule

Reject the null hypothesis if the test statistics $F > F_{(.05, 1,8)} = 5.32$

Test Statistics

$F = 72.73$ (From ANOVA table)

Conclusion

Reject the null hypothesis. With 0.05 level of significance, we claim that there is a linear association between the number of times a carton is transferred and the number of broken ampules.

Example 5

Obtain the T statistic for the test in part (b) and demonstrate numerically its equivalence to the F statistic obtained in part (b).

Hypotheses

Null Hypothesis $H_0: \beta_1 = 0$

Alternative Hypothesis $H_1: \beta_1 \neq 0$

Decision Rule

Reject the null hypothesis if the test statistics $|T| > t_{(.025, 8)} = 2.3065$ (Note: $2.3065^2 = 5.32$)

Test Statistics $|T| = (4 - 0)/.469 = 8.529$ (Note: $8.529^2 = 72.73$)

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	10.20000	0.66332	15.38	<.0001	8.67037	11.72963
X	shipment route	1	4.00000	0.46904	8.53	<.0001	2.91839	5.08161

Example 5

Calculate R^2 and r. What proportion of the variation in Y is accounted for by introducing X into the regression model?

R^2 = SSR/SST = 160/177.6 = .9009

r = .9492

The proportion of the variation in Y being accounted for by introducing X into the regression model is 90.09%.