Inference in Simple Linear Regression

1. Inference in Regression Models

The parameters β_1 and β_0 in the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ are estimated as:

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

Just like other statistics (such as the sample mean or variance) the estimates b_1 and b_0 are functions of the observed values Y_i, which are functions of the random errors ε_i. Thus b_1 and b_0 are themselves random variables and b_1 and b_0 each has a probability distribution, called the *sampling distribution*. The *sampling distribution* of b_1 (respectively b_0) refers to "the different values of b_1 (b_0) that would be obtained with repeated sampling when the levels of the independent variables X are held constant from sample to sample" (textbook, pg. 41).

Statistical inference concerning population parameters such as β_1 and β_0 consists in testing hypotheses and constructing confidence intervals for that parameter. Inference concerning a parameter is based on the sampling distribution of that parameter.

For (simple or multiple) linear regression models statistical inference is commonly carried out for

1. the regression coefficients β_1 and β_0 (hypothesis tests and confidence intervals)
2. confidence intervals for the regression line (i.e., “where do I think the population regression line lies?”)
3. prediction intervals for individual observations (i.e., “where do I think a single new observation will fall?”)
4. F test of the significance of the regression model as a whole (hypothesis test only)

1. Sampling Distribution of b_1

The sampling distribution of b_1 can be standardized by

$$\frac{b_1 - \beta_1}{\sigma\{b_1\}} \sim N(0,1)$$

where $\sigma\{b_1\}$ is the standard error of b_1; however, this is not known and estimated by $s\{b_1\}$. So,

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n - 2)$$

2. Inference on β_1 and β_0

Table 1. Formulas for Inference on b_1 and b_0

Slope b_1	
Estimated standard error of b_1	$s\{b_1\} = \sqrt{\dfrac{MSE}{\sum (X_i - \bar{X})^2}}$
Estimated sampling distribution of b_1	$\dfrac{b_1 - \beta_1}{s\{b_1\}} \sim t(n-2)$
Confidence limits for CI on β_1	$b_1 \pm t(1 - \alpha; n - 2)s\{b_1\}$
Test statistic for H_0: $\beta_1 = \beta_1^0$	$t^* = \dfrac{b_1 - \beta_1^0}{s\{b_1\}}$
Intercept b_0	
Estimated standard error of b_0	$s\{b_0\} = \sqrt{MSE\left(\dfrac{1}{n} + \dfrac{\bar{X}^2}{\sum (X_i - \bar{X})^2}\right)}$
Estimated sampling distribution of b_0	$\dfrac{b_0 - \beta_0}{s\{b_0\}} \sim t(n-2)$
Confidence limits for CI on β_0	$b_0 \pm t(1 - \alpha; n - 2)s\{b_0\}$
Test statistic for H_0: $\beta_0 = \beta_0^0$	$t^* = \dfrac{b_0 - \beta_0^0}{s\{b_0\}}$

Note: β_1^0 and β_0^0 denote hypothetical values of the parameters
When $\beta_1^0 = 0$ (the most common type of hypothesis) then $t^* = b_1/s\{b_1\}$
Recall: $MSE = \Sigma e_i^2/(n - 2)$

The estimated *standard error* $s\{b_1\}$ is typically provided in the standard regression printout (labeled Std Error in Table 2).

2. Inference on β_1

We look at inference on β_1 first because it is the most common.

Confidence Interval for β_1

From Table 1 the confidence interval the CI for β_1 is

$$b_1 \pm t(1 - \alpha; n - 2)s\{b_1\}$$

Ph.D. Example: find the 95% CI for β_1, the coefficient of publications in the simple regression of time since Ph.D.

$b_1 = 1.9830$

$$s\{b_1\} = \sqrt{\frac{MSE}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{117.0396}{293.3333}} = 0.6317$$

Choose $\alpha = .05$; then $t(1 - \alpha/2; n - 2) = t(0.975; 13) = 2.16$ (from statistical program or table)

Therefore the 95% CI for β_1 is

Lower bound of CI = 1.9830 - (2.16)(0.6317) = 0.619
Upper bound of CI = 1.9830 + (2.16)(0.6317) = 3.347

The 95% CI is [0.619, 3.347]. Over repeated sampling, 95 out of 100 confidence intervals will contain β_1. We are 95% confident that this interval contains β_1.

Two-sided hypothesis test for β_1

Example: Test the hypothesis that the coefficient of time $\beta_1 = 0$. The setup is

Step 1: Set up null and alternative hypothesis

H_0: $\beta_1 = 0$ ("null hypothesis")
H_1: $\beta_1 \neq 0$ ("alternative hypothesis")

Step 2: Choose a significance level

$\alpha = .05$

Step 3: Calculate test statistic

$t^* = (b_1 - 0)/s\{b_1\} = b_1/s\{b_1\} = 1.9830/.6317 = 3.139$

Step 4: Determine the p-value or the critical value

P-value approach: Find the 2-tailed p-value

```
data pvalue;
tobs = 3.139;
df = 13;
prob = 2*(1-probt(tobs, df));
run;
```

p-value = .0078

Critical value approach: Determine the critical value

With $\alpha = .05$, $t(1 - \alpha/2; n - 2)$ is $t(0.975; 13) = 2.16$

Step 5: Make a decision

P-value approach:
$p \leq \alpha$ Reject H_0
$p > \alpha$ Fail to reject H_0

Critical value approach
if $|t^*| > t(1 - \alpha/2; n - 2)$, reject H_0
if $|t^*| \leq t(1 - \alpha/2; n - 2)$, fail to reject H_0

Since $|t^*| = 3.13 > 2.16$ or $p < .05$, reject H_0 and conclude H_1 ($\beta_1 \neq 0$) at the .05 level.

One-sided test for β_1

Hint: It is often easier to write down H_1 (the "alternative hypothesis") first; then H_0 is the *complement* of H_1, i.e.

H_0: $\beta_1 \leq 0$
H_1: $\beta_1 > 0$

To carry out the test

- if b_1 is in a direction opposite to H_1 (i.e., if $b_1 > 0$), then there is no point in doing the test and H_1 can be rejected at the outset
- otherwise (b_1 is in a direction compatible with H_1) using the P-value approach one simply calculates the 1-sided P-value associated with b_1 by *dividing the 2-tailed P-value (shown on the regression printout) by 2*

In the example the 2-sided P-value is .0078; thus the 1-sided P-value is (.0078)/2 = .004
Since P-value = .004 < .05 = α, conclude H_1: $\beta_1 > 0$.

Comparing Two-sided and One-sided Tests for β_1

Comparing the two types of tests it appears that *the 1-sided test is "easier" (i.e., more likely to turn up significant) than the corresponding 2-sided test*. (For example, the P-value of the 1-sided test is half the P-value of the 2-sided test.)

Thus there is an incentive to use 1-sided tests to increase the chance of significant results. It is considered legitimate to use a 1-sided test *whenever one has a genuine directional hypothesis concerning β_1*. This opinion on the use of one-tailed tests is widely shared by reviewers of professional journals. However, some statisticians recommend using 2-sided tests exclusively, on the ground that the 2-sided test is conservative.

3. Inference on β_0

CIs and tests for β_0 are carried out in exactly the same way as for β_1.
Q - Using information from the printout

- calculate the 95% CI for β_0
- test the 2-sided hypothesis that $\beta_0 \neq 0$

3. Inference for Mean Response $E\{Y_h\}$

We recognize that these predicted Y values are subject to error and we often wish to determine a CI around a predicted value to reflect the degree of precision or uncertainty in a prediction.

A critical point to understand is that there are two kinds of confidence intervals that one can construct around predicted scores:

1. The first arises from the fact that from one sample to another the slope of the regression line will vary due to sampling variability. As a result, predicted Y values associated with a given X value will vary from one sample to another.
2. A second kind of interval estimate is associated with accuracy of predictions.

The second kind of CI is called a prediction interval and will be discussed in the next section.

Let X_h = the level of X we wish to estimate the mean response (does not necessarily correspond to a data point X_i)

1. Sampling Distribution of $\hat{Y}_h$

The mean response $E\{Y_h\}$ is estimated as $\hat{Y}_h = b_0 + b_1X_h$. Thus the variance of the sampling distribution of $\hat{Y}_h$ is affected by variance in both b_0 and b_1 sampling and by how far X_h is from the sample mean of X. The way in which the variance of $\hat{Y}_h$ depends on the distance of X_h from $X_.$ is shown in the figure to the right: given a change in b_1, the change in $\hat{Y}_h$ is larger further away from the mean.

FIGURE 2.3 Effect on $\hat{Y}_h$ of Variation in b_1 from Sample to Sample in Two Samples with Same Means $\bar{Y}$ and $\bar{X}$.

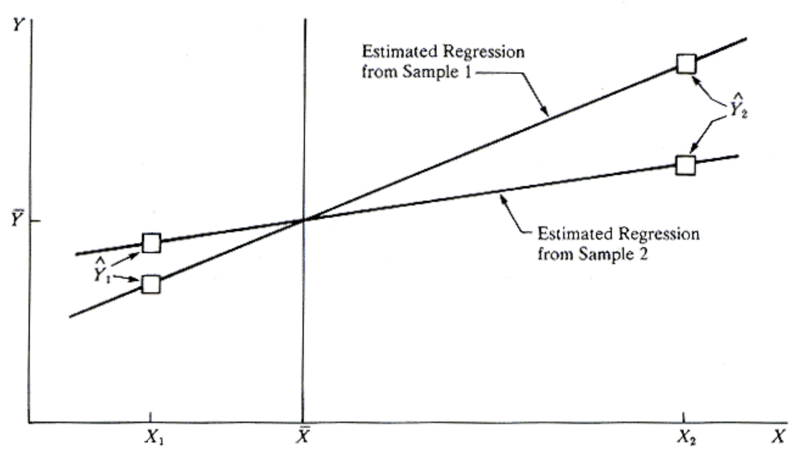

Table 2. Formulas for Inference on $\hat{Y}_h$

Point estimator of E{Y_h}	$\hat{Y}_h = b_0 + b_1 X_h$
Estimated standard error of $\hat{Y}_h$	$s\{\hat{Y}_h\} = \sqrt{MSE\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right)}$
Estimated sampling distribution of $\hat{Y}_h$	$\frac{\hat{Y}_h - E\{Y_h\}}{s\{\hat{Y}_h\}} \sim t(n-2)$
Confidence limits for CI on E{Y_h}	$\hat{Y}_h \pm t(1 - \alpha; n-2)s\{\hat{Y}_h\}$
Test statistic	Not often used

- An important feature of the standard error of $\hat{Y}_h$ is that it will become larger as the deviation of X_i from the mean of X increases. In other words, predictions of Y become less stable for individuals farther from the mean of X.
- The CI for E{Y_h} reflects the degree of variability in a predicted Y value across repeated sampling. These CIs will become wider for individuals whose X scores are farther from the mean of X.

2. CI for Mean Response E{Y_h}

Given the regression of the number of publications on time since Ph.D., calculate an interval estimate for E{Y_h} when time is 5, i.e. $X_h = 5$.

$$\hat{Y}_h = 4.7310 + 1.9830(5) = 14.646$$

$$s\{\hat{Y}_h\} = \sqrt{117.0396\left(\frac{1}{15} + \frac{7.1111}{293.3333}\right)} = 3.262$$

Choose $\alpha = .05$; then $t(1 - \alpha/2; n - 2) = t(0.975; 13) = 2.16$ (from statistical program or table)

Therefore the 95% CI for E{Y_h} is

Lower bound of CI = 14.646 - (2.16)(3.262) = 7.600
Upper bound of CI = 14.646 + (2.16)(3.262) = 21.692

The 95% CI is [7.600, 21.692]. Over repeated sampling, 95 out of 100 confidence intervals will contain E{$Y_h|X_h = 5$}. We are 95% confident that this interval contains E{Y_h}.

3. CI for the Entire Regression Line - the Working-Hotelling Confidence Band

The Working-Hotelling confidence band is a CI for the entire regression line $E\{Y\} = \beta_0+\beta_1 X$. See SAS code.

4. Prediction Interval for a New Observation $Y_{h(new)}$

1. Sampling Distribution of $Y_{h(new)}$

One estimates $Y_{h(new)}$ as $\hat{Y}_h = b_0 + b_1 X_h$, the value of $\hat{Y}$ on the regression line corresponding to X_h.

Variation in $Y_{h(new)}$ is affected by two sources:

- variation in $\hat{Y}_h$, the estimated mean of the distribution of Y given X_h, namely $\sigma^2\{\hat{Y}_h\}$
- variation in the probability distribution of Y around its mean given X_h, namely σ^2

Therefore,
$\sigma^2\{Y_{h(new)}\} = \sigma^2 + \sigma^2\{\hat{Y}_h\}$

Table 3. Formulas for Inference on $Y_{h(new)}$

Point estimator of $Y_{h(new)}$	$\hat{Y}_h = b_0 + b_1 X_h$
Estimated standard error of $Y_{h(new)}$	$s\{pred\}=\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right)}$
Estimated sampling distribution of $Y_{h(new)}$	$\frac{\hat{Y}_{h(new)} - E\{Y_h\}}{s\{pred\}} \sim t(n-2)$
Confidence limits for CI on $Y_{h(new)}$	$\hat{Y}_h \pm t(1 - \propto; n - 2)s\{pred\}$
Test statistic	Not often used

2. CI for $Y_{h(new)}$

Given the regression of the number of publications on time since Ph.D., calculate an interval estimate for $Y_{h(new)}$ when time is 5, i.e. $X_h = 5$.

$\hat{Y}_h = 4.7310 + 1.9830(5) = 14.646$

$s\{\hat{Y}_h\} = \sqrt{117.0396\left(1 + \frac{1}{15} + \frac{7.1111}{293.3333}\right)} = 11.3$

Choose $\alpha = .05$; then $t(1-\alpha/2; n-2) = t(0.975; 13) = 2.16$ (from statistical program or from table)
Therefore the 95% CI for $E\{Y_h\}$ is

Lower bound of CI = 14.646 - (2.16)(11.3) = -9.805
Upper bound of CI = 14.646 + (2.16)(11.3) = 39.097

The 95% CI is [-9.805, 39.097]. Over repeated sampling, 95 out of 100 confidence intervals will contain $Y_{h(new)}$. We are 95% confident that this interval contains $Y_{h(new)}$.

Note that the 95% CI for $Y_{h(new)}$ is *considerably* wider than the CI for the mean response $E\{Y_h\}$

5. F Test for Entire Regression (Alternative Test of $\beta_1 = 0$)

1. F Test of $\beta_1 = 0$

Here this test is overkill, since the t test can be used to test the hypothesis that $\beta_1 = 0$.
But the F test generalizes to multiple regression to test the hypothesis that *all the regression coefficients are zero*.

2. Partitioning Sum of Squares Total

The total variation of Y_i from the sample mean of Y, $(Y_i - \bar{Y})$ can be decomposed into two components:

$Y_i - \bar{Y}$	=	$\hat{Y}_i - \bar{Y}$	+	$Y_i - \hat{Y}_i$
(total deviation of Y_i from mean)		(deviation of predicted value from mean)		(deviation of Y_i from predicted value)

Next take the sum of the squares of each deviation over all observations in the sample.

Sum of squares:	$\Sigma(Y_i - \bar{Y})^2$	$\Sigma(\hat{Y}_i - \bar{Y})^2$	$\Sigma(Y_i - \hat{Y}_i)^2$
Name:	SSTO for *sum of squares total*	SSR for *sum of squares regression*	SSE for *sum of squares error*
Meaning:	(total variation in Y)	(variation in Y accounted for by regression line)	(variation in Y around regression line)

SSE is also called *residual* sum of squares. The basic ANOVA result (or *theorem*) is that the sums of squared deviations stand in the same relation as the (unsquared) deviations, so that:

$$
\begin{array}{lcccccc}
 & \Sigma(Y_i - \bar{Y})^2 & = & \Sigma(\hat{Y}_i - \bar{Y})^2 & + & \Sigma(Y_i - \hat{Y}_i)^2 \\
\text{or} & \text{SSTO} & = & \text{SSR} & + & \text{SSE}
\end{array}
$$

3. Partitioning of Degrees of Freedom

To each sum of squares correspond degrees of freedom (df). Degrees of freedom are additive.

$$
\begin{array}{ccccc}
n - 1 & = & 1 & + & (n - 2) \\
\text{df for SSTO} & & \text{df for SSR} & & \text{df for SSE}
\end{array}
$$

4. Expected Mean Squares

The ANOVA table breaks down the total sum of squares and associated degrees of freedom along with mean squares. Recall MSE is an estimate of the *variance of the residuals* σ^2.

Source of Variation	**SS**	**df**	**MS**	**F**
Regression	$SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$	1	MSR = SSR/1	F* = MSR/MSE
Error	$SSE = \Sigma(Y_i - \hat{Y}_i)^2$	n – 2	MSE = SSE/(n-2)	
Total	$SSTO = \Sigma(Y_i - \bar{Y})^2$	n – 1		

One can show that
$E\{MSE\} = \sigma^2$
$E\{MSR\} = \sigma^2 + \beta_1^2\ \Sigma(X_i - \bar{X})^2$

Note that if $\beta_1=0$, $E\{MSR\} = E\{MSE\} = \sigma^2$.

Hypotheses for the F Test:

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

Test statistic: F* = MSR/MSE

If $\beta_1 = 0$, $E\{F^*\} = \sigma^2/\sigma^2 = 1$.
If $\beta_1 \neq 0$, $E\{F^*\} > 1$.

Thus the larger F*, the more likely that $\beta_1 \neq 0$.

$F^* = MSR/MSE \sim F(1; n - 2)$

3. Carrying Out the F Test

Example: Carry out the F test for the regression of the number of publications on time since Ph.D.

Step 1: Set up null and alternative hypothesis

H_0: $\beta_1 = 0$ ("null hypothesis")

H_1: $\beta_1 \neq 0$ ("alternative hypothesis")

Step 2: Choose a significance level

$\alpha = .05$

Step 3: Calculate test statistic

$F^* = 1153.41/117.04 = 9.85$

Step 4: Determine the p-value or the critical value

P-value approach: Find the 2-tailed p-value

```
data pvalue;
Fobs = 9.85;
ndf = 1;
ddf = 13;
prob = 1-probf(fobs,ndf,ddf);
run;
```

p-value = .0078

Critical value approach: Determine the critical value

With $\alpha = .05$, $F(1 - \alpha ;1, n - 2) = F(0.95;1,13) = 4.67$

Step 5: Make a decision

P-value approach:

$p \leq \alpha$ Reject H_0

$p > \alpha$ Fail to reject H_0

Critical value approach

if $F^* > F(1 - \alpha; 1, n - 2)$, reject H_0

if $F^* \leq F(1 - \alpha; 1, n - 2)$, fail to reject H_0

Since $F^* = 9.85 > 4.67$ or $p < .05$, reject H_0 and conclude H_1 ($\beta_1 \neq 0$) at the .05 level.

4. Equivalence of t Test and F Test

In the simple linear regression model, $F^* = (t^*)^2$.

Example: In the Ph.D. example, the squared t-ratio for b_1 is equal to F^*, i.e.

$(t^*)^2 = (3.139)^2 = 9.85 = F^*$

This is no longer true in the multiple regression model.