Chapter 2 – Inferences in Regression Analysis

Review (Appendix A): Rules Concerning Linear Functions of Random Variables (P. 645)

Let $Y_1,\ldots,Y_n$ be *n* random variables. Consider the function $\sum_{i=1}^{n} a_i Y_i$ where the coefficients $a_1,\ldots,a_n$ are constants. Then, we have:

$$E\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i E\{Y_i\}$$

$$\sigma^2\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \sigma\{Y_i, Y_j\}$$

When $Y_1,\ldots,Y_n$ are independent (as in the model in Chapter 1), the variance of the linear combination simplifies to:

$$\sigma^2\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i^2 \sigma^2\{Y_i\}$$

When $Y_1,\ldots,Y_n$ are independent, the covariance of two linear functions $\sum_{i=1}^{n} a_i Y_i$ and $\sum_{i=1}^{n} c_i Y_i$ can be written as:

$$\sigma\left\{\sum_{i=1}^{n} a_i Y_i, \sum_{i=1}^{n} c_i Y_i,\right\} = \sum_{i=1}^{n} a_i c_i \sigma^2\{Y_i\}$$

We will use these rules to obtain the distribution of the estimators $b_0, b_1, \hat{Y} = b_0 + b_1 X$

Throughout this chapter, we assume that the normal error regression model is applicable.

Objectives:

- Confidence intervals for regression parameters
- Significance tests for regression parameters
- Confidence interval for a mean response
- Prediction interval for a future observation
- Analysis of variance (ANOVA) approach to regression
- General linear test approach
- Coefficient of determination: descriptive measure of association

2.1 Inferences Concerning β_1

Confidence interval for a parameter has this format:

estimate ± critical value × SE(estimate)

Recall that the least squares estimate of the slope parameter, b_1 , is a linear function of the observed responses $Y_1,\ldots,Y_n$:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}Y_i = \sum_{i=1}^{n} k_i Y_i$$

$$k_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Note that $E\{Y_i\} = \beta_0 + \beta_1 X_i$, so that the expected value of b_1 is:

$$E\{b_1\} = \sum_{i=1}^{n} k_i E\{Y_i\} = \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}(\beta_0 + \beta_1 X_i)$$

$$= \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\left\{\beta_0\sum_{i=1}^{n}(X_i - \overline{X}) + \beta_1\sum_{i=1}^{n}(X_i - \overline{X})X_i\right\}$$

Note that $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$ (why?), so that the first term in the brackets is 0, and that we can add $\beta_1 \overline{X}\sum_{i=1}^{n}(X_i - \overline{X}) = 0$ to the last term to get:

$$E\{b_1\} = \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\left\{\beta_1\sum_{i=1}^{n}(X_i - \overline{X})X_i - \beta_1\sum_{i=1}^{n}(X_i - \overline{X})\overline{X}\right\} = \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\beta_1\sum_{i=1}^{n}(X_i - \overline{X})^2 = \beta_1$$

Thus, b_1 is an unbiased estimator of the parameter β_1.

To obtain the variance of b_1, recall that $\sigma^2\{Y_i\} = \sigma^2$. Thus:

$$\sigma^2\{b_1\} = \sum_{i=1}^{n} k_i^2\sigma^2\{Y_i\} = \sum_{i=1}^{n}\left[\frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]^2\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}\sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Note that the variance of b_1 decreases when we have larger sample sizes (as long as the added X levels are not placed at the sample mean $\overline{X}$). Since σ^2 is unknown in practice, and must be estimated from the data, we obtain the estimated variance of the estimator b_1 by replacing the unknown σ^2 with its unbiased estimate $s^2 = MSE$:

$$s^2\{b_1\} = \frac{s^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{MSE}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

with estimated standard error:

$$s\{b_1\} = \frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}} = \frac{\sqrt{MSE}}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

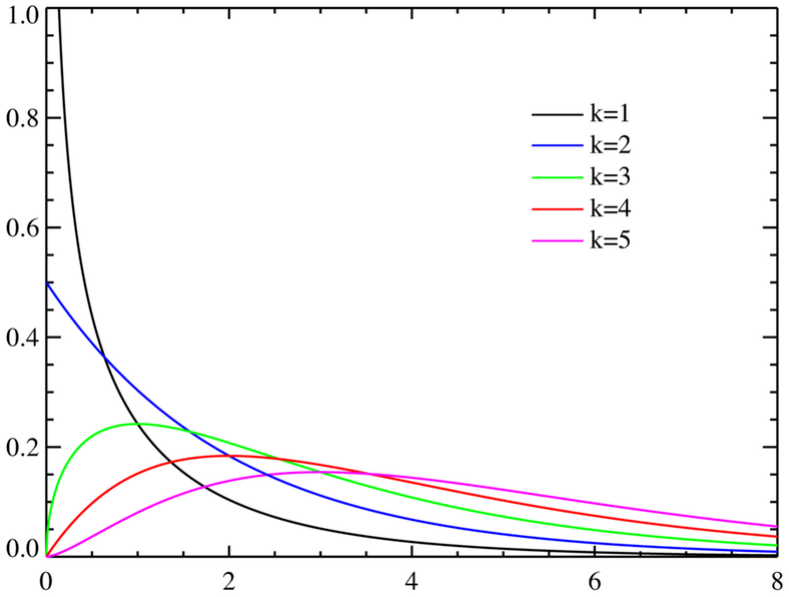
Further, **the sampling distribution of b_1 is normal, that is:**

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$$

since under the current model, b_1 is a linear function of independent, normal random variables $Y_1, \ldots, Y_n$.

Theorem: For the normal error regression model, SSE/σ^2 is distributed as χ^2 with n-2 degrees of freedom and is independent of b_0 and b_1.

Here is the graph for the probability density curves of χ^2 distribution with k degrees of freedom.



Making use of this theory, we obtain the following result that allows us to make inferences concerning β_1:

$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n-2)$ **where *t*(*n*-2) represents Student's t-distribution with *n*-2 degrees of freedom.**

Proof:

(1) $\frac{b_1-\beta_1}{s\{b_1\}}=\frac{b_1-\beta_1}{\sigma\{b_1\}}\div\frac{s\{b_1\}}{\sigma\{b_1\}}$,

(2) $\frac{b_1-\beta_1}{\sigma\{b_1\}}$ is a standard normal variable,

(3) $\frac{s^2\{b_1\}}{\sigma^2\{b_1\}}=\frac{MSE}{\sigma^2}=\frac{SSE/(n-2)}{\sigma^2}=\frac{SSE}{\sigma^2(n-2)}\sim\frac{\chi^2(n-2)}{(n-2)}$

(4) $\frac{z}{\sqrt{\chi^2_{df}/df}}=t(df)$

Confidence Interval for β_1

As a result of the fact that $\frac{b_1-\beta_1}{s\{b_1\}}\sim t(n-2)$, we obtain the following probability statement:

$P\{t(\alpha/2;n-2)\le\frac{b_1-\beta_1}{s\{b_1\}}\le t(1-\alpha/2;n-2)\}\quad=\quad 1-\alpha$ where $t(\alpha/2;n-2)$ is the $(\alpha/2)100^{th}$ percentile of the *t*-distribution with *n*-2 degrees of freedom. Note that since the *t*-distribution is symmetric around 0, we have that $t(\alpha/2;n-2)=-t(1-\alpha/2;n-2)$. Traditionally, we obtain the table values corresponding to $t(1-\alpha/2;n-2)$, which is the value of that leaves an upper tail area of $\alpha/2$. The following algebra results in obtaining a $(1-\alpha)100\%$ confidence interval for β_1:

$$P\{t(\alpha/2;n-2)\le\frac{b_1-\beta_1}{s\{b_1\}}\le t(1-\alpha/2;n-2)\}$$
$$=P\{-t(1-\alpha/2;n-2)\le\frac{b_1-\beta_1}{s\{b_1\}}\le t(1-\alpha/2;n-2)\}$$
$$=P\{-t(1-\alpha/2;n-2)s\{b_1\}\le b_1-\beta_1\le t(1-\alpha/2;n-2)s\{b_1\}\}$$
$$=P\{-b_1-t(1-\alpha/2;n-2)s\{b_1\}\le-\beta_1\le-b_1+t(1-\alpha/2;n-2)s\{b_1\}\}$$
$$=P\{b_1+t(1-\alpha/2;n-2)s\{b_1\}\ge\beta_1\ge b_1-t(1-\alpha/2;n-2)s\{b_1\}\}$$

This leads to the following rule for a $(1-\alpha)100\%$ confidence interval for β_1:

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

Some statistical software packages print this out automatically (e.g. EXCEL and SPSS). Other packages simply print out estimates and standard errors only (e.g. SAS).

Tests Concerning β_1

We can also make use of the fact that $\frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{n-2}$ to test hypotheses concerning the slope parameter. As with means and proportions (and differences of means and proportions), we can conduct one-sided and two-sided tests, depending on whether a priori a specific directional belief is held regarding the slope. More often than not (but not necessarily), the null value for β_1 is 0 (the mean of Y is independent of X) and the alternative is that β_1 is positive (1-sided), negative (1-sided), or different from 0 (2-sided). The alternative hypothesis must be selected before observing the data.

2-sided tests

- Null Hypothesis: $H_0 : \beta_1 = \beta_{10}$
- Alternative (Research Hypothesis): $H_A : \beta_1 \neq \beta_{10}$
- Test Statistic: $t^* = \frac{b_1 - \beta_{10}}{s\{b_1\}}$
- Decision Rule: Conclude H_A if $|t^*| \geq t(1-\alpha/2; n-2)$, otherwise conclude H_0
- P-value: $2P(t(n-2) > |t^*|)$

All statistical software packages (to my knowledge) will print out the test statistic and P-value corresponding to a 2-sided test with β_{10}=0.

1-sided tests (Upper Tail)

- Null Hypothesis: $H_0 : \beta_1 = \beta_{10}$
- Alternative (Research Hypothesis): $H_A : \beta_1 > \beta_{10}$
- Test Statistic: $t^* = \frac{b_1 - \beta_{10}}{s\{b_1\}}$
- Decision Rule: Conclude H_A if $t^* \geq t(1-\alpha; n-2)$, otherwise conclude H_0
- P-value: $P(t(n-2) > t^*)$

A test for positive association between Y and X (H_A:β_1>0) can be obtained from standard statisical software by first checking that b_1 (and thus t^*) is positive, and cutting the printed P-value in half.

1-sided tests (Lower Tail)

- Null Hypothesis: $H_0 : \beta_1 = \beta_{10}$
- Alternative (Research Hypothesis): $H_A : \beta_1 < \beta_{10}$
- Test Statistic: $t^* = \dfrac{b_1 - \beta_{10}}{s\{b_1\}}$
- Decision Rule: Conclude H_A if $t^* \leq -t(1-\alpha; n-2)$, otherwise conclude H_0
- *P*-value: $P(t(n-2) < t^*)$

A test for negative association between Y and X (H_A:β_1<0) can be obtained from standard statisical software by first checking that b_1 (and thus t^*) is negative, and cutting the printed *P*-value in half.

2.2 Inferences Concerning β_0

Recall that the least squares estimate of the intercept parameter, b_0 , is a linear function of the observed responses $Y_1, \ldots, Y_n$:

$$b_0 = \overline{Y} - b_1 \overline{X} = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(X_i - \overline{X})\overline{X}}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right] Y_i = \sum_{i=1}^{n} l_i Y_i$$

Recalling that $E\{Y_i\} = \beta_0 + \beta_1 X_i$:

$$E\{b_0\} = \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(X_i - \overline{X})\overline{X}}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right] (\beta_0 + \beta_1 X_i) \quad = \quad \beta_0 \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(X_i - \overline{X})\overline{X}}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right] + \beta_1 \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(X_i - \overline{X})\overline{X}}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right] X_i$$

$$= \beta_0 (1-0) + \beta_1 \left[\frac{1}{n} \sum_{i=1}^{n} X_i - \overline{X} \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right] = \beta_0 + \beta_1 (\overline{X} - \overline{X}(1)) = \beta_0$$

Thus, b_0 is an unbiased estimator or the parameter β_0. Below, we obtain the variance of the estimator of b_0.

$$\sigma^2\{b_0\} = \sum_{i=1}^{n}\left[\frac{1}{n} - \frac{(X_i-\overline{X})\overline{X}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]^2\sigma^2 = \sigma^2\sum_{i=1}^{n}\left[\frac{1}{n^2} + \frac{\overline{X}^2(X_i-\overline{X})^2}{\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right)^2} - \frac{2\overline{X}(X_i-\overline{X})}{n\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]$$

$$= \sigma^2\left[\frac{n}{n^2} + \frac{\overline{X}^2}{\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right)^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 - \frac{2\overline{X}}{n\sum_{i=1}^{n}(X_i-\overline{X})^2}\sum_{i=1}^{n}(X_i-\overline{X})\right]$$

$$= \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]$$

Note that the variance will decrease as the sample size increases, as long as X values are not all placed at the mean. Further, the sampling distribution is normal under the assumptions of the model. The estimated standard error of b_0 replaces σ^2 with its unbiased estimate $s^2=MSE$ and taking the square root of the variance.

$$s\{b_0\} = s\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}} = \sqrt{MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]}$$

Note that $\frac{b_0-\beta_0}{s\{b_0\}} \sim t(n-2)$, allowing for inferences concerning the intercept parameter β_0 when it is meaningful, namely when X=0 is within the range of observed data.

Confidence Interval for β_0

$b_0 \pm t(1-\alpha/2; n-2)s\{b_0\}$

Example: Look at the example at the end of Chapter 1 again.

R codes

```
# To have 95% confidence intervals for beta0 and beta1
confint(fit, level=0.95)
```

It is also useful to obtain **the covariance of b_0 and b_1,** as they are only independent under very rare circumstances:

$$\sigma\{b_0, b_1\} = \sigma\left\{\sum_{i=1}^{n} l_i Y_i, \sum_{i=1}^{n} k_i Y_i\right\} = \sum_{i=1}^{n} l_i k_i \sigma^2\{Y_i\}$$

$$= \sum_{i=1}^{n}\left[\frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]\frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sigma^2$$

$$= \frac{\sigma^2}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}\sum_{i=1}^{n}(X_i - \overline{X}) - \frac{\sigma^2 \overline{X}}{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$= 0 - \frac{\sigma^2 \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad = \quad -\frac{\sigma^2 \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

In practice, $\overline{X}$ is usually positive, so that the intercept and slope estimators are usually negatively correlated. We will use the result shortly.

2.3 Considerations on Making Inferences Concerning β_0 and β_1

Normality of Error Terms

If the data are approximately normal, simulation results have shown that using the *t*-distribution will provide approximately correct significance levels and confidence coefficients for tests and confidence intervals, respectively. Even if the distribution of the errors (and thus *Y*) is far from normal, in large samples the sampling distributions of b_0 and b_1 have sampling distributions that are approximately normal as results of central limit theorems. This is sometimes referred to as *asymptotic normality*.

Interpretations of Confidence Coefficients and Error Probabilities

Since *X* levels are treated as fixed constants, these refer to the case where we repeated the experiment many times at the current set of *X* levels in this data set. In this sense, it's easier to interpret these terms in controlled experiments where the experimenter has set the levels of *X* (such as time and temperature in a laboratory type setting) as opposed to observational studies, where nature determines the *X* levels, and we may not be able to reproduce the same conditions repeatedly. This will be covered later.

Spacing of *X* Levels

The variances of b_0 and b_1 (for given n and σ^2) decrease as the X levels are more spread out, since their variances are inversely related to $\sum_{i=1}^{n}(X_i - \overline{X})^2$. However, there are reasons to choose a diverse range of X levels for assessing model fit. This is covered in Chapter 4.

Power of Tests

The **power** of a statistical test refers to the probability that we reject the null hypothesis when the null hypothesis is false, which is 1 minus the probability of a Type II error (π=1-β), where π denotes the power of the test and β is the probability of a Type II error (failing to reject the null hypothesis when the alternative hypothesis is true). The following procedure can be used to obtain the power of the test concerning the slope parameter with a 2-sided alternative.

1) Write out null and alternative hypotheses: $H_0 : \beta_1 = \beta_{10}$ $\quad$ $H_A : \beta_1 \neq \beta_{10}$

2) Obtain the noncentrality measure, the standardized distance between the true value of β_1 and the value under the null hypothesis (β_{10}): $\delta = \dfrac{|\beta_1 - \beta_{10}|}{\sigma\{b_1\}}$

3) Choose the probability of a Type I error (α=0.05 or α=0.01)

4) Determine the degrees of freedom for error: $df = n$-2

5) Refer to Table B.5 (pages 671-672), identifying α (page), δ (row) and error degrees of freedom (column). The table provides the power of the test under these parameter values.

Note that the power increases within each tables as the noncentrality measure increases for a given degrees of freedom, and as the degrees of freedom increases for a given noncentrality measure.

2.4 Confidence Interval for $E\{Y_h\}=\beta_0+\beta_1 X_h$

When we wish to estimate the mean at a hypothetical X value (within the range of observed X values), we can use the fitted equation at that value of $X=X_h$ as a **point estimate**, but we have to include the uncertainty in the regression estimators to construct a confidence interval for the mean.

Parameter: $E\{Y_h\} = \beta_0 + \beta_1 X_h$

Estimator: $\hat{Y}_h = b_0 + b_1 X_h$

We can obtain the variance of the estimator (as a function of $X=X_h$) as follows:

$$\sigma^2\left\{\hat{Y}_h\right\} = \sigma^2\{b_0 + b_1 X_h\} = \sigma^2\{b_0\} + X_h^2\sigma^2\{b_1\} + 2X_h\sigma\{b_0, b_1\}$$

$$= \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right] + X_h^2\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} + 2X_h\left[-\frac{\sigma^2\overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$$

$$= \sigma^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$$

Estimated standard error of estimator: $s\{\hat{Y}_h\} = \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]}$

$\frac{\hat{Y}_h - E\{Y_h\}}{s\{\hat{Y}_h\}} \sim t(n-2)$ which can be used to construct confidence intervals for the mean response at specific X levels, and tests concerning the mean (tests are rarely conducted).

(1-α)100% Confidence Interval for $E\{Y_h\}$:

$$\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{\hat{Y}_h\}$$

2.5 Predicting a Future Observation When *X* is Known

If β_0, β_1, σ were known, we'd know that the distribution of responses when $X=X_h$ is normal with mean $\beta_0 + \beta_1 X_h$ and standard deviation σ. Thus, making use of the normal distribution (and equivalently, the empirical rule) we know that if we took a sample item from this distribution, it is very likely that the value fall within 2 standard deviations of the mean. That is, we would know that the probability that the sampled item lies within the range $(\beta_0 + \beta_1 X_h - 2\sigma, \beta_0 + \beta_1 X_h + 2\sigma)$ is approximately 0.95.

In practice, we don't know the mean $\beta_0 + \beta_1 X_h$ or the standard deviation σ. However, we just constructed a (1-α)100% Confidence Interval for $E\{Y_h\}$, and we have an estimate of σ (s). Intuitively, we can approximately use the logic of the previous paragraph (with the estimate of σ) across the range of believable values for the mean. Then our prediction interval spans the lower tail of the normal curve centered at the lower bound for the mean to the upper tail of the normal curve centered at the upper bound for the mean. See Figure 2.5 on page 64 of the text book.

The prediction error for the new observation is the difference between the observed value and its predicted value: $Y_h - \hat{Y}_h$. Since the data are assumed to be independent, the new (future) value is independent of its predicted value, since it wasn't used in the regression analysis. The variance of the prediction error can be obtained as follows:

$$\sigma^2\{pred\} = \sigma^2\{Y_h - \hat{Y}_h\} = \sigma^2\{Y_h\} + \sigma^2\{\hat{Y}_h\} = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$$

$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$$

It is the sum of the (1) variance of the distribution of Y at $X=X_h$ namely, σ^2, and (2) the variance of the sampling distribution $\hat{Y}_h$, namely, $\sigma^2\{\hat{Y}_h\}$. An unbiased estimator is:

$$s^2\{pred\} = MSE\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$$

(1-α)100% Prediction Interval for New Observation When $X=X_h$

$$\hat{Y}_h \pm t(\alpha/2; n-2)\sqrt{MSE\left[1+\frac{1}{n}+\frac{(X_h-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]}$$

It is a simple extension to obtain a prediction for the mean of *m* new observations when $X=X_h$. The sample mean of *m* observations is $\frac{\sigma^2}{m}$ and we get the following variance for for the error in the prediction mean:

$$s^2\{predmean\} = MSE\left[\frac{1}{m}+\frac{1}{n}+\frac{(X_h-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]$$

and the obvious adjustment to the prediction interval for a single observation.

(1-α)100% Prediction Interval for the Mean of *m* New Observations When $X=X_h$

$$\hat{Y}_h \pm t(\alpha/2; n-2)\sqrt{MSE\left[\frac{1}{m}+\frac{1}{n}+\frac{(X_h-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]}$$

2.6 Confidence Band for the Entire Regression Line (Working-Hotelling Method)

$$\hat{Y}_h \pm Ws\{\hat{Y}_h\} \qquad W = \sqrt{2F(1-\alpha;2,n-2)}$$

Notice that the formula for the boundary values is of exactly the same form as for the confidence limits for the mean response at X_h, except that the t multiple has been replaced by the W multiple.

The confidence band is wider than the confidence limits for the mean response at a given X level, because the confidence band encompass the entire regression line and one is able to draw conclusion about any values of X.

Toluca Company Example (Page 19 of the text, data set CH01TA01.txt)

The Toluca Company manufactures refrigeration equipment as well as many replacement parts. In the past, one of the replacement parts has been produced periodically in lots of varying sizes. When a cost improvement program was taken, company officials wished to determine the optimum lot size for producing this part. One key input is to ascertain the optimum lot size was the relationship between size and labor hours required to produce the lot. **CH01TA01.txt** ontains the data on lot size and work hours for 25 recent production runs.

(1) Find a 95% confidence interval for $E\{Y_h\}$ when the lot size $X_h = 65,\ 100$.

(2) Suppose that the next lot to be produced consists of $X_h = 100$ units, find a 95% prediction interval.

R codes

```
# Read in the data

# (1) File CH01PR19.thourst is saved under the working directory
Data <- read.table(file="CH01PR19.txt", header=FALSE)

# (2) or give the full path like below
Data <- read.table(file="C:/Hongmei/CH01TA01.txt", header=FALSE)

# To get size and hours separately, you can use
size <- Data[,1]
hours <- Data[,2]

# Scatter plot
plot(hours ~ size, xlab="size", ylab="hours")

# fit the simple linear regression line
fit <- lm(hours ~ size)
fit

# add the regression line
abline(fit)

# New observations
new <- data.frame(size=c(65, 100))

# Confidence interval of the mean response
predict(fit, new, interval="confidence", level=0.95, se.fit=TRUE)

#prediction limits for new observation
predict(fit, new, interval="prediction")
```

R Output

```
> fit
Call:
lm(formula = hours ~ size)
```

```
Coefficients:
(Intercept)         size
      62.37         3.57

> # Estimation of mean response for new observations
> new <- data.frame(size=c(65, 100))
> # Confidence interval of the mean response
> predict(fit, new, interval="confidence", level=0.95,
se.fit=TRUE)
$fit
       fit      lwr      upr
1 294.4290 273.9129 314.9451
2 419.3861 389.8615 448.9106

> #prediction limits for new observation
> predict(fit, new, interval="prediction")
       fit      lwr      upr
1 294.4290 191.3676 397.4904
2 419.3861 314.1604 524.6117
```

2.7 Analysis of Variance Approach to Regression

Consider the total deviations of the observed responses from the mean: $Y_i - \overline{Y}$. When these terms are all squared and summed up, this is referred to as the **total sum of squares (SSTO)**.

$$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

The more spread out the observed data are, the larger SSTO will be.

Now consider the deviation of the observed responses from their fitted values based on the regression model: $Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i) = e_i$. When these terms are squared and summed up, this is referred to as the **error sum of squares (SSE)**. We've already encounterd this quantity and used it to estimate the error variance.

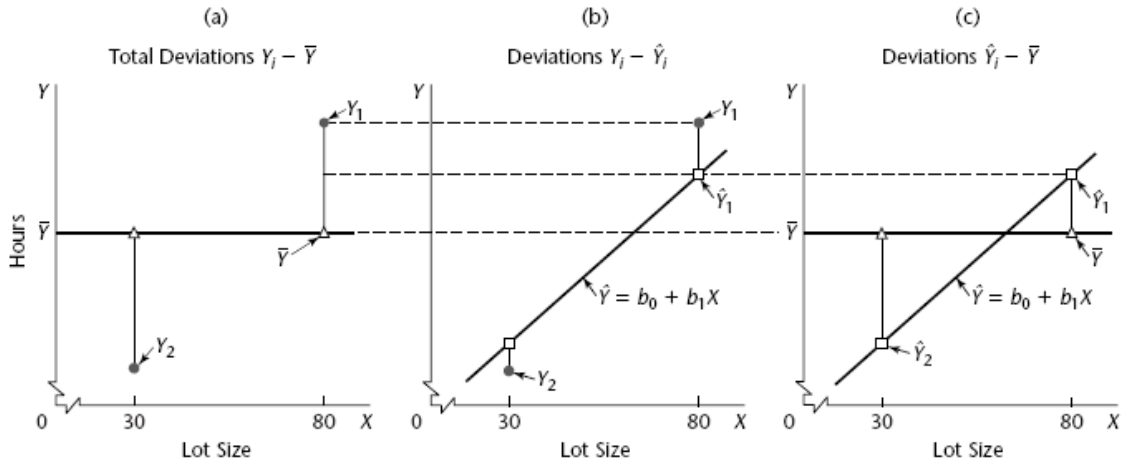
$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

When the observed responses fall close to the regression line, SSE will be small. When the data are not near the line, SSE will be large.

Finally, there is a third quantity, representing the deviations of the predicted values from the mean. Then these deviations are squared and summed up, this is referred to as the **regression sum of squares** (**SSR)**.

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

FIGURE 2.7 **Illustration of Partitioning of Total Deviations $Y_i - \bar{Y}$—Toluca Company Example (not drawn to scale; only observations Y_1 and Y_2 are shown).**



The error and regression sums of squares sum to the total sum of squares:
$SSTO = SSR + SSE$ which can be seen as follows:

$$Y_i - \bar{Y} = Y_i - \bar{Y} + \hat{Y}_i - \hat{Y}_i = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \quad \Rightarrow$$

$$(Y_i - \bar{Y})^2 = [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2 = (Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \quad \Rightarrow$$

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} \left[(Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \right] =$$

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) =$$

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n} e_i (b_0 + b_1 X_i - \bar{Y}) =$$

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + 2\left[b_0 \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} e_i X_i - \bar{Y} \sum_{i=1}^{n} e_i \right] =$$

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + 2(0) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = SSE + SSR$$

The last term was 0 since $\sum e_i = \sum e_i X_i = 0$.

$SSTO = SSR + SSE$ **says the total sums of squares can be partitioned into (1) model (explained by regression) and (2) error (unexplained/residuals).**

Each sum of squares has associated with **degrees of freedom**. The total degrees of freedom is $df_T = n\text{-}1$. The error degrees of freedom is $df_E = n\text{-}2$. The regression degrees of freedom is $df_R = 1$. Note that the error and regression degrees of freedom sum to the total degrees of freedom: $n-1 = 1+(n-2)$.

Mean squares are the sums of squares divided by their degrees of freedom:

$$MSR = \frac{SSR}{1} \qquad MSE = \frac{SSE}{n-2}$$

Note that *MSE* was our estimate of the error variance, and that we don't compute a total mean square. It can be shown that the expected values of the mean squares are:

$$E\{MSE\} = \sigma^2 \qquad E\{MSR\} = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Note that these expected mean squares are the same if and only if β_1=0.

The Analysis of Variance is reported in tabular form:

Source	**df**	**SS**	**MS**	**F**
Regression	1	SSR	MSR=SSR/1	F=MSR/MSE
Error	*n*-2	SSE	MSE=SSE/(*n*-2)	
C Total	*n*-1	SSTO		

***F* Test of $\beta_1 = 0$ versus $\beta_1 \neq 0$**

As a result of Cochran's Theorem (stated on page 69-70 of text book), we have a test of whether the dependent variable *Y* is linearly related to the predictor variable *X*. This is a very specific case of the *t*-test described previously. Its full utility will be seen when we consider multiple predictors. The test proceeds as follows:

- Null hypothesis: $H_0 : \beta_1 = 0$
- Alternative (Research) Hypothesis: $H_A : \beta_1 \neq 0$
- Test Statistic: $TS : F^* = \frac{MSR}{MSE}$
- Rejection Region: $RR : F^* \geq F(1-\alpha; 1, n-2)$
- *P*-value: $P\{F(1, n-2) \geq F^*\}$

Critical values of the *F*-distribution (indexed by numerator and denominator degrees' of freedom) are given in Table B.4, pages 665-670.

Note that this is a very specific version of the *t*-test regarding the slope parameter, specifically a 2-sided test of whether the slope is 0. Mathematically, the tests are identical:

$$t^* = \frac{b_1 - 0}{s\{b_1\}} = \frac{\frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}}{\sqrt{\frac{MSE}{\sum (X_i - \overline{X})^2}}} = \frac{\frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}}}{\sqrt{MSE}}$$

Note that:

$$MSR = SSR = \sum (\hat{Y}_i - \overline{Y})^2 = \sum (b_0 + b_1 X_i - \overline{Y})^2 =$$

$$nb_0^2 + b_1^2 \sum X_i^2 + n\overline{Y}^2 + 2b_0 b_1 \sum X_i - 2nb_0 \overline{Y} - 2b_1 \overline{Y} \sum X_i \quad =$$

$$n(\overline{Y} - b_1 \overline{X})^2 + b_1^2 \sum X_i^2 + n\overline{Y}^2 + 2(\overline{Y} - b_1 \overline{X}) b_1 n \overline{X} - 2n(\overline{Y} - b_1 \overline{X})\overline{Y} - 2b_1 \overline{Y} n \overline{X} \quad =$$

$$n\overline{Y}^2 + nb_1^2 \overline{X}^2 - 2nb_1 \overline{X}\overline{Y} + b_1^2 \sum X_i^2 + n\overline{Y}^2 + 2nb_1 \overline{X}\overline{Y} - 2nb_1^2 \overline{X}^2 - 2n\overline{Y}^2 + 2nb_1 \overline{X}\overline{Y} - 2nb_1 \overline{X}\overline{Y}$$

$$= (n\overline{Y}^2 + n\overline{Y}^2 - 2n\overline{Y}^2) + (-2nb_1 \overline{X}\overline{Y} + 2nb_1 \overline{X}\overline{Y} + 2nb_1 \overline{X}\overline{Y} - 2nb_1 \overline{X}\overline{Y}) + (b_1^2 \sum X_i^2 + nb_1^2 \overline{X}^2 - 2nb_1^2 \overline{X}^2)$$

$$= 0 + 0 + b_1^2 \sum X_i^2 - nb_1^2 \overline{X}^2 = b_1^2 \sum (X_i - \overline{X})^2 = \left[\frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \right]^2 \sum (X_i - \overline{X})^2$$

$$= \frac{\left[\sum (X_i - \overline{X})(Y_i - \overline{Y}) \right]^2}{\sum (X_i - \overline{X})^2}$$

Thus:

$$(t^*)^2 = \left[\frac{\frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}}}{\sqrt{MSE}} \right]^2 = \frac{\left[\sum (X_i - \overline{X})(Y_i - \overline{Y}) \right]^2}{\sum (X_i - \overline{X})^2} = \frac{MSR}{MSE} = F^*$$

Further, the critical values are equivalent: $(t(1-\alpha/2; n-2))^2 = F(1-\alpha; 1, n-2)$, check this from the two tables. Thus, the tests are equivalent.

Take a look at the ANOVA for Toluca Company Example (Page 19 of the text, data set CH01TA01.txt) again.

```
> anova(fit)
Analysis of Variance Table

Response: hours
          Df Sum Sq Mean Sq F value    Pr(>F)
size       1 252378  252378  105.88 4.449e-10 ***
Residuals 23  54825    2384
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

2.8 General Linear Test Approach

This is a very general method of testing hypotheses concerning regression models. We first consider the simple linear regression model, and testing whether *Y* is linearly associated with *X*. We wish to test $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$.

Full Model

This is the model specified under the alternative hypothesis, also referrred to as the unrestricted model. Under simple linear regression with normal errors, we have:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Using least squares (and maximum likelihood) to estimate the model parameters ($\hat{Y}_i = b_0 + b_1 X_i$), we obtain the error sum of squares for the full model:

$$SSE(F) = \sum (Y_i - (b_0 + b_1 X_i))^2 = \sum (Y_i - \hat{Y}_i)^2 = SSE$$

Reduced Model

This the model specified by the null hypothesis, also referred to as the restricted model. Under simple linear regression with normal errors, we have:

$$Y_i = \beta_0 + 0 X_i + \varepsilon_i = \beta_0 + \varepsilon_i$$

Under least squares (and maximum likelihood) to estimate the model parameter, we obtain $\overline{Y}$ as the estimate of β_0, and have $b_0 = \overline{Y}$ as the fitted value for each observation. We when get the following error sum of squares under the reduced model:

$$SSE(R) = \sum (Y_i - b_0)^2 = \sum (Y_i - \overline{Y})^2 = SSTO$$

Test Statistic

The error sum of squares for the full model will always be less that or equal to the error sum of squares for reduced model, by definition of least squares. The test statistic will be:

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}$$ where df_R, df_F are the error degrees of freedom for the full and reduced models. We will use this method throughout course.

For the simple linear regression model, we obtain the following quantities:

$$SSE(F) = SSE \qquad df_F = n-2 \qquad SSE(R) = SSTO \qquad df_R = n-1$$

thus the *F*-Statistic for the General Linear Test can be written:

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} = \frac{\frac{SSTO - SSE}{(n-1)-(n-2)}}{\frac{SSE}{n-2}} = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{MSR}{MSE}$$

Thus, for this particular null hypothesis, the general linear test "generalizes" to the *F*-test.

2.9 Descriptive Measures of Association

Along with the slope, *Y*-intercept, and error variance; several other measures are often reported.

Coefficient of Determination (R^2)

The coefficient of determination measures the proportion of the variation in *Y* that is "explained" by the regression on *X*. It is computed as the regression sum of squares divided by the total (corrected) sum of squares. Values near 0 imply that the regression model has done little to "explain" variation in *Y*, while values near 1 imply that the model has "explained" a large portion of the variation in *Y*. If all the data fall exactly on the fitted line, $R^2=1$. The coefficient of determination will lie beween 0 and 1.

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \qquad 0 \le r^2 \le 1$$

Coefficient of Correlation (*r*)

The coefficient of correlation is a measure of the strength of the linear association between *Y* and *X*. It will always be the same sign as the slope estimate (b_1), but it has several advantages:

- In some applications, we cannot identify a clear dependent and independent variable, we just wish to determine how two variables vary together in a population (peoples heights and weights, closing stock prices of two firms, etc). Unlike the slope estimate, the coefficient of correlation does not depend on which variable is labeled as *Y*, and which is labeled as *X*.
- The slope estimate depends on the units of *X* and *Y*, while the correlation coefficient does not.
- The slope estimate has no bound on its range of potential values. The correlation coefficient is bounded by –1 and +1, with higher values (in absolute value) implying stronger linear association (it is not useful in measuring nonlinear association which may exist, however).

$$r = \text{sgn}(b_1)\sqrt{R^2} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})(Y_i - \overline{Y})}} = \frac{s_x}{s_y} b_1 \qquad -1 \le r \le 1$$

where sgn(b_1) is the sign (positive or negative) of b_1, and s_x, s_y are the sample standard deviations of *X* and *Y*, respectively.

2.10 Issues in Applying Regression Analysis

- When using regression to predict the future, the assumption is that the conditions are the same in future as they are now. Clearly any future predictions of economic variables such as tourism made prior to September 11, 2001 would not be valid.
- Often when we predict in the future, we must also predict *X*, as well as *Y*, especially when we aren't controlling the levels of *X*. Prediction intervals using methods described previously will be too narrow (that is, they will overstate confidence levels).
- Inferences should be made only within the range of *X* values used in the regression analysis. We have no means of knowing whether a linear association continues outside the range observed. That is, we should not **extrapolate** outside the range of *X* levels observed in experiment.
- Even if we determine that *X* and *Y* are associated based on the *t*-test and/or *F*-test, we cannot conclude that changes in *X* **cause** changes in *Y*. Finding an association is only one step in demonstrating a causal relationship.
- When multiple tests and/or confidence intervals are being made, we must adjust our confidence levels. This is covered in Chapter 4.
- When X_i is a random variable, and not being controlled, all methods described thus far hold, as long as the X_i are independent, and their probability distribution does not depend on $\beta_0, \beta_1, \sigma^2$.

Example using R: GPA Problem (Data: CH01PR19.txt; Question 2.4, 2.13).

1.19. **Grade point average.** 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). The results of the study follow. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	118	119	120
X_i:	21	14	28	...	28	16	28
Y_i:	3.897	3.885	3.778	...	3.914	1.860	2.948

a. Obtain the least squares estimates of β_0 and β_1, and state the estimated regression function.

b. Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?

c. Obtain a point estimate of the mean freshman GPA for students with ACT test score $X = 30$.

d. What is the point estimate of the change in the mean response when the entrance test score increases by one point?

2.4. Refer to **Grade point average** Problem 1.19.

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

b. Test, using the test statistic t^*, 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. State the alternatives, decision rule, and conclusion.

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

2.13. Refer to **Grade point average** Problem 1.19.

a. Obtain a 95 percent interval estimate of the mean freshman GPA for students whose ACT test score is 28. Interpret your confidence interval.

b. Mary Jones obtained a score of 28 on the entrance test. Predict her freshman GPA using a 95 percent prediction interval. Interpret your prediction interval.

c. Is the prediction interval in part (b) wider than the confidence interval in part (a)? Should it be?

d. Determine the boundary values of the 95 percent confidence band for the regression line when $X_h = 28$. Is your confidence band wider at this point than the confidence interval in part (a)? Should it be?