Inference in Regression Analysis

- Inference Concerning $\beta_0$ and $\beta_1$

- Estimate $E\{Y_h\}$ and Predict New Observations

- Confidence Band for Regression Line

- Analysis of Variance Approach to Linear Regression

- Association Between $X$ and $Y$ in Regression Model
Regression Model Assumptions

Throughout the lecture we will assume the normal error regression model:

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n, \quad (1) \]

where:

- \( \beta_0 \) and \( \beta_1 \) are parameters
- \( X_i \) are known constants
- \( \epsilon_i \) are independent \( \mathcal{N}(0, \sigma^2) \).

If \( \epsilon_i \) are not independent, the \( Y_i \) are correlated. E.g. the autocorrelation in time series data or repeated measurements in longitudinal data analysis. The least squares method is not appropriate for these data, the maximum likelihood principle can be used to model the dependence.
If $\epsilon_i$ don’t have constant variance $\sigma^2$, the model has heteroscedastic errors. In this case weighted regression methods could be employed (it can be derived through maximum likelihood principle).

In applications $X_i$ could also be random. E.g. in the measurement error model, the $X_i$ are assume to be measured with errors.
Inference Concerning $\beta_0$ and $\beta_1$

- Point Estimation: $b_0$ and $b_1$
- Confidence Intervals
- Hypothesis Testing
- Some Considerations in Inference
**Interpretation of $\beta_1$ and $\beta_0$**

$\beta_1$ is the slope of the regression line. It indicates the change in $E(Y)$ per unit increase in $X$.

If $\beta_1 = 0$, $Y_i = \beta_0 + \epsilon_i$. There is no linear association between $Y$ and $X$ and the means of $Y$ are all equal. For the normal error regression model (1), when $\beta_1 = 0$ the probability distribution of $Y$ are identical. So there is no relation of any type between $Y$ and $X$.

$\beta_0$ is the intercept of the regression line. When the scope of the model includes $X = 0$, $\beta_0$ gives $E(Y|X = 0)$, the mean of the probability distribution of $Y$ at $X = 0$. When the scope of the model doesn’t cover $X = 0$, $\beta_0$ doesn’t have any particular meaning as a separate term in the regression model.
Point Estimation of $\beta_0$ and $\beta_1$

\[ b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \quad b_0 = \bar{Y} - b_1 \bar{X}. \]  \hspace{1cm} (2)

As we discussed in the last lecture, the $b_1$ and $b_0$ have normal distribution:

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

\[ b_0 \sim N \left( \beta_0, \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right\} \right) \]  \hspace{1cm} (3)

The distribution of $b_0$ and $b_1$ refers to their different values that would be
obtained with repeated sampling when the levels of the predictor variable $X$ are held constant from sample to sample.

Or you can think of $b_1$ and $b_0$ as functions of random variables $Y_i$ in (2), where $X_i$ are fixed constants. So they are also random variables and have distributions.

For $b_1$ we have the Gauss-Markov theorem

$b_1$ has minimum variance among all unbiased linear estimators of the form: $\hat{\beta}_1 = c_i Y_i$.

Why? $\iff \text{Cov}(\hat{\beta}_1 - b_1, b_1) = 0$

Note: read the textbook on the detailed mathematical deductions of the mean and variance of $b_0$ and $b_1$. 
Confidence Intervals for $\beta_0$ and $\beta_1$

In the distribution of $b_0$ and $b_1$ (3), there is an extra unknown parameter $\sigma^2$ except $\beta_0$ and $\beta_1$. We estimate $\sigma^2$ by the unbiased MSE:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}, \quad (n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-2). \quad (4)$$

By replacing $\sigma$ with $\hat{\sigma}$ in (3), we have the unbiased estimators for $Var(b_1)$ and $Var(b_0)$:

$$s^2(b_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \quad s^2(b_0) = \hat{\sigma}^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right\}. \quad (5)$$
Also we can prove that

\[ Cov(b_1, \hat{\sigma}^2) = Cov(b_0, \hat{\sigma}^2) = 0, \]  

and hence

\[ \frac{(b_j - \beta_j)}{s(b_j)} \sim t(n - 2), j = 0, 1. \]  

So we can make the following probability statements:

\[
\Pr \left\{ t(\alpha/2, n - 2) \leq \frac{b_j - \beta_j}{s(b_j)} \leq t(1 - \alpha/2, n - 2) \right\} = 1 - \alpha, \ j = 0, 1. \tag{8} \]

\( t \)-distribution is symmetric around 0: \( t(\alpha/2, n - 2) = -t(1 - \alpha/2, n - 2). \)

So we have the confidence limits for \( \beta_j \):

\[ b_j \pm t(1 - \alpha/2, n - 2)s(b_j), \ j = 0, 1. \tag{9} \]
Hypothesis Testing: Two-sided Test

Test the following hypotheses

\[ H_0 : \beta_j = \beta_j^0, \quad H_a : \beta_j \neq \beta_j^0, \quad j = 0, 1. \]

Test statistic:

\[ t_j^* = \frac{b_j - \beta_j^0}{s(b_j)}, \quad t_j^*|H_0 \sim t(n - 2). \]

Rejection rules under Type-I error \( \alpha \):

- If \( |t_j^*| \leq t(1 - \alpha/2, n - 2) \), conclude \( H_0 \)
- If \( |t_j^*| > t(1 - \alpha/2, n - 2) \), conclude \( H_a \).
P-value:

\[ p = \Pr(|T| > |t^*_j|) = 2F(-|t^*_j|) = 2F\left(-\frac{|b_j - \beta_0^j|}{s(b_j)}\right), \quad (10) \]

where \( F(\cdot) \) is the distribution function of \( t(n - 2) \). Notice that \( F(\cdot) \) is a fixed function and P-value is a decreasing function of \( |t^*_j| \). The P-value, as a function of \( b_j \), is also a random variable and has distributions. We have the following important result

**Under \( H_0 \), the P-value (10) is uniformly distributed in interval \([0, 1]\).**

Sketchy proof:

\[
\Pr(p \leq \alpha) = \Pr\left\{F(-|t^*_j|) \leq \alpha/2\right\} = \Pr\left\{|t^*_j| \geq -t(\alpha/2, n - 2)\right\} \\
= 2 - 2F\left\{-t(\alpha/2, n - 2)\right\} = \alpha.
\]
If we use P-value as our test statistic, the rejection rules become

If $p > \alpha$, conclude $H_0$. If $p \leq \alpha$, conclude $H_a$.

Because the P-value is uniformly distributed in interval $[0, 1]$ under $H_0$, it is obvious that the rejection rule has Type-I error equal to $\alpha$.

So p-value as a test statistic has a simple rejection rule: reject $H_0$ if $p < \alpha$, when we call the test significant. This explains why sometimes we call the P-value the observed level of significance.
Hypothesis Testing: One-sided Test

Test the following hypotheses

\[ H_0 : \beta_j \leq \beta_j^0, \quad H_a : \beta_j > \beta_j^0, \quad j = 0, 1. \]

Test statistic:

\[ t_j^* = \frac{b_j - \beta_j^0}{s(b_j)}, \quad t_j^*\big|_{H_0} \sim t(n - 2). \]

Rejection rules under Type-I error \( \alpha \):

\[
\begin{align*}
&\text{If } t_j^* \leq t(1 - \alpha, n - 2), \text{ conclude } H_0 \\
&\text{If } t_j^* > t(1 - \alpha, n - 2), \text{ conclude } H_a.
\end{align*}
\]
P-value:

\[ p = \Pr(T > t^*_j) = 1 - F(t^*_j) = 1 - F \left( \frac{b_j - \beta^0_j}{s(b_j)} \right). \]  \hspace{1cm} (11)

It is easy to show that this P-value also has a uniform distribution in interval \([0, 1]\). So we also has the simple rejection rule: reject \( H_0 \) if \( p < \alpha \), when we call the test significant.
Some Considerations on Making Inference on $\beta_0$ and $\beta_1$

- Departure of the $Y$ Distribution from Normality
- Interpretation of Confidence Intervals
- Spacing of the $X$ Levels
- Test Power
Departure of the $Y$ Distribution from Normality

If not seriously, previous results hold approximately. When $Y$ are distributed far from normal, $b_0$ and $b_1$ have the property of asymptotic normality - their distributions approach normality under very general conditions as the sample size increases. So with sufficiently large samples, previous confidence intervals and decision rules still apply. And for large sample, $t$ is replaced by standard normal distribution. Why? roughly $\bar{\epsilon} = \sum_{j=1}^{n} \epsilon_j/n \to N(0, \sigma^2/n)$. 
Interpretation of Confidence Intervals

In the regression model (1), $X_i$ are assumed as known constants, the confidence intervals are interpreted with respect to taking repeated samples in which the $X$ observations are kept at the same levels as in the observed sample. So basically only $\epsilon_i$ are changing across each random sample. The $1 - \alpha$ confidence intervals for $\beta_j$ are interpreted to mean that:

If many independent samples are taken where the levels of $X$ are the same as the observed values $(X_1, \cdots, X_n)$, and an $(1 - \alpha)$ confidence interval is constructed for each sample, $100 \times (1 - \alpha)$ of the intervals will contain the true value of $\beta_j$.

So the randomness in the confidence interval really comes from the error part $\epsilon$. $\beta_j$ are fixed constants. There are no randomness associated with their
values, only our estimates are random. Each time for different sample, our confidence interval will be different.
Spacing of the $X$ Levels

The variance formula (3) for estimates $b_0$ and $b_1$ indicates that for given $n$ and $\sigma^2$, they are affected by the spacing of the $X$ levels in the observed data: the greater is the spread in the $X$ levels, the larger is the quantity $\sum_{i=1}^{n}(X_i - \bar{X})^2$ and the smaller is the variance of $b_1$. 
Test Power

The power of a test is the probability that the decision rule will lead to conclusion \( H_a \) when \( H_a \) in fact holds:

\[
\text{Power} = \Pr(\text{reject } H_a | H_a \text{ is true}) \tag{12}
\]

For previous two-sided test, the power is given by

\[
\Pr(|t^*_j| > t(1 - \alpha/2, n - 2) | \delta),
\]

and for the one-sided test, the power is given by

\[
\Pr(t^*_j > t(1 - \alpha, n - 2) | \delta),
\]
where $\delta$ is the noncentrality measure, i.e. a measure of how far the true value of $\beta_j$ are from $\beta_0^j$:

$$\delta = \frac{\beta_j - \beta_0^j}{\sigma(b_j)}.$$ 

So to calculate power, we have to know the true values for all the parameters: $\beta_0$, $\beta_1$ and $\sigma$. 
Example: Toluca Company Data

\[ Y = 62.37 + 3.57X \]

\[ Y = 70 + 3X \]
The point estimates:

\[ \hat{\beta}_0 = b_0 = 62.37, \hat{\beta}_1 = b_1 = 3.57, \hat{\sigma} = \sqrt{MSE} = 48.8. \]

If the LotSize can be zero, the number 62.37 gives the estimated average WorkHours for zero LotSize. Otherwise there is no meaning for \( b_0 = 62.37 \). \( b_1 = 3.57 \) gives the estimated average increase of WorkHours per unit increase of LotSize. The \( \hat{\sigma} = 48.8 \) gives the estimated variation of WorkHours at each observed LotSize levels

Variance estimates for \( b_j \):

\[ s(b_0) = 26.18, \quad s(b_1) = 0.35. \]
95% confidence intervals

$$\beta_0 : (8.2, 116.5), \ \beta_1 : (2.9, 4.3).$$

Test

$$H_0 : \beta_1 = 0, \ H_a : \beta_1 \neq 0.$$ 

Test statistic and P-value

$$t^*_1 = \frac{|b_1|}{s(b_1)} = 10.29, \ p = 4.45 \times 10^{-10}.$$ 

while $$t(1 - 0.05/2, n - 2) = 2.07,$$ we conclude that $$\beta_1 \neq 0.$$ So there is a linear relationship between WorkHours and LotSize and the rest result is highly significant.

Test

$$H_0 : \beta_0 = 0, \ H_a : \beta_0 \neq 0.$$
Test statistic and P-value

\[ t_0^* = \frac{|b_0|}{s(b_0)} = 2.38, \ p = 0.026. \]

while \( t(1 - 0.05/2, n - 2) = 2.07 \), so we conclude that the intercept \( \beta_0 \neq 0 \).

Power calculation: suppose that \( \beta_0 = 62, \beta_1 = 3, \sigma = 40 \). For test

\[ H_0 : \beta_1 = 0, \ H_a : \beta_1 \neq 0, \]

the power is

\[ \Pr(|t| > 2.07) \approx 1. \]

For test

\[ H_0 : \beta_0 = 0, \ H_a : \beta_0 \neq 0, \]

the power is

\[ \Pr(|t| > 2.07) = 0.79. \]
Estimate $E(Y_h)$ and Predict New Observations

- $E(Y_h)$: point estimation and confidence interval
- Prediction: single observation $Y_h$ and multiple samples
Inference about $E(Y_h)$

Under our normal error regression model (1), the mean response when $X = X_h$, $E(Y_h) = \beta_0 + X_h\beta_1$, can be estimated by

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

which is a linear function of $Y_i$, and

$$\hat{Y}_h \sim N \left( E(Y_h), \sigma^2 \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right\} \right).$$
Substitute \( \sigma \) with \( MSE \), we estimate the variance of \( \hat{Y}_h \) by

\[
s^2(\hat{Y}_h) = MSE \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right\}, \tag{15}
\]

and

\[
\frac{\hat{Y}_h - E(Y_h)}{s(\hat{Y}_h)} \sim t(n - 2), \tag{16}
\]

which can be used to derive the \( 1 - \alpha \) confidence interval for \( E(Y_h) \)

\[
\hat{Y}_h \pm t(1 - \alpha/2, n - 2)s(\hat{Y}_h). \tag{17}
\]

Note that in the variance formula (15), the further from \( \bar{X} \) is \( X_h \), the greater is the quantity \( (X_h - \bar{X})^2 \) and the larger is the variance of \( \hat{Y}_h \).
Toluca Company Data

- Regression Line
- 95% CI for E(Y)

LotSize vs. WorkHours
Predict New Observation $Y_h$

The new observation $Y_h$ corresponding to a given level $X_h$ is a new sample and independent of all the observed samples $(Y_1, \cdots, Y_n)$. And we assume the underlying regression model still applies to new observations.

Under normal error regression model (1)

$$Y_h = \beta_0 + X_h \beta_1 + \epsilon_h \sim N(\beta_0 + X_h \beta_1, \sigma^2), \epsilon_h \perp Y_i, i = 1, \cdots, n,$$

where $\epsilon_h$ is the extra term different from $E(Y_h)$. We can “estimate” $\epsilon_h$ by its mean (zero), $\beta_0, \beta_1$ by $b_0, b_1$, and we obtain the following point estimate

$$\hat{Y}_h = b_0 + X_h b_1.$$
Due to the independence of $\epsilon_h$ and $Y_i$, we have $\epsilon_h \perp (b_0, b_1)$ and

$$Y_h - \hat{Y}_h = (\beta_0 - b_0) + X_h(\beta_1 - b_1) + \epsilon_h \sim N\left(0, \sigma^2_{pred}\right), \quad (19)$$

where (refer to (14))

$$\sigma^2_{pred} = \text{Var}(Y_h - \hat{Y}_h) = \text{Var}(\hat{Y}_h) + \text{Var}(\epsilon_h)$$

$$= \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right\}. \quad (20)$$

Substitute $\sigma^2$ with $MSE$ we obtain the following prediction variance estimation

$$s^2(\text{pred}) = MSE \left\{ 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right\}, \quad (21)$$
and
\[ \frac{Y_h - \hat{Y}_h}{s(\text{pred})} \sim t(n - 2). \]  

Thus we obtain the \(1 - \alpha\) “prediction limits” for the new observation
\[ \hat{Y}_h \pm t(1 - \alpha/2, n - 2)s(\text{pred}). \]
Toluca Company Data

- Regression Line
- 95% CI for E(Y)
- 95% PI for Y

LotSize vs. WorkHours graph showing the relationship between lot size and work hours.
Comparison of Inference on $E(Y_h)$ and $Y_h$

- Compared to (17), the confidence interval for $E(Y_h)$, the only difference is the variance estimation:

$$s(pred) > s(\hat{Y}_h),$$

where the extra variation comes from the randomness of $\epsilon_h$, which is the unavoidable variation within the probability distribution of $Y|X$.

- The prediction limits (23), unlike the confidence interval (17) for a mean response $E(Y_h)$, are sensitive ro departure from normal error assumption. The reason is that when we conclude $Y_h - \hat{Y}_h$ is normal distribution, we heavily rely on the distribution of single error term $\epsilon_h$. 
• The confidence coefficient $1 - \alpha$ for the prediction limits (23) refers to the taking of repeated samples based on the same set of $X$ values, and calculating prediction limits for $Y_h$ for each sample.

• Confidence interval represents an inference on a parameter and is an interval that is intended to cover the value of the parameter. While prediction interval, on the other hand, is a statement about the possible value of a random variable.
Predict Mean of $m$ New Observations for Given $X_h$

Based on our normal error regression model assumption (1), we can write down the mean (random variable) as

$$\bar{Y}_h = \beta_0 + X_h \beta_1 + \sum_{j=1}^{m} \epsilon_j / m,$$

where

$$\bar{\epsilon}_m = \sum_{j=1}^{m} \epsilon_j / m \sim N(0, \sigma^2 / m), \quad \bar{\epsilon}_m \perp Y_i, i = 1, \cdots, n.$$

Similarly we can derive

$$\bar{Y}_h - \hat{Y}_h \sim N(0, \sigma_{predmean}^2), \quad (24)$$
where

\[
\sigma^2_{\text{predmean}} = \text{Var}(\bar{Y}_h) + \text{Var}(\hat{Y}_h) = \sigma^2 \left\{ \frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right\},
\]

and it is estimated by

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

And similar \(1 - \alpha\) prediction limits can be derived

\[
\hat{Y}_h \pm t(1 - \alpha/2, n - 2)s(\text{predmean}).
\]
Confidence Band for Regression Line

The confidence band for the entire regression line \( E(Y) = \beta_0 + X_h \beta_1 \) identifies the region in which the entire line lies, which is useful for assessing the appropriateness of a fitted regression function.

The Working-Hotelling \( 1 - \alpha \) confidence band for the regression line for normal error model has the following forms

\[
\hat{Y} \pm W \times s(\hat{Y}_h), \quad W = \sqrt{2F(1 - \alpha, 2, n - 2)}.
\]  \( (25) \)

The confidence band (25) is wider than the confidence interval (17), which only apply at the single level \( X_h \), while confidence band apply to all possible \( X \) levels.
Toluca Company Data

- Regression Line
- 95% CI for E(Y)
- 95% WH-CB for E(Y)
- 95% PI for Y

LotSize vs. WorkHours graph.
Comparison of Interval Length

\[
\alpha
\]

\[
W
\]

\[
\sqrt{2F(1 - \alpha, 2, 23)}
\]

\[
t(1 - \alpha/2, 23)
\]
Note that

- The confidence band (25) applies to the entire regression line over all values of \( X \) from \(-\infty\) to \( \infty \). The confidence band indicates the proportion of the time that the estimating procedure will yield a band that covers the line, in a long series of samples in which \( X \) observations are kept at the same as the observed values. So the band is random, but the underlying line is some fixed values.

- In applications, the confidence band is ignored for those part of no interest. And the confidence coefficient for a limited line segment is higher than \( 1 - \alpha \), which serves as a lower bound to the actual confidence coefficient.

Toluca Company Data

- Regression Line
- 95% Exact-CB
- 95% CI for E(Y)
- 95% WH-CB for E(Y)
Analysis of Variance Approach to Regression Analysis

- Decomposition of Variance and Degree of Freedom
- F-test and Equivalence to T-test
### Variance Decomposition

\[
Y_i - \bar{Y} = Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y} = e_i + b_1(X_i - \bar{X}),
\]

where \(e_i\) are the residuals and \(\hat{Y}_i = b_0 + X_i b_1 = \bar{Y} + b_1(X_i - \bar{X})\).

Since

\[
\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} e_i X_i = 0,
\]

we have the following decomposition of variance

\[
\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2.
\]
$SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is called the *total sum of squares* and measures the uncertainty of $Y$ without considering any other variables.

$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ is just the *error sum of squares* and reflects the uncertainty of $Y$ after utilizing $X$ predictor.

$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ is called the *regression sum of squares*. It accounts for the difference between $SSTO$ and $SSE$ and measures part of the variation of $Y$ which is associated with the regression line.
Decomposition of Degree of Freedom

As discussed earlier, $SSE$ has $n - 2$ degree of freedom, and we have $n - 1$ degree of freedom associated with $SSTO$. $SSR$ has one degree of freedom.

The error mean square, $MSE = \frac{SSE}{n-2}$, is an unbiased estimator for $\sigma^2$. We can similarly define regression mean square, $MSR = \frac{SSR}{1}$. We have

$$E(MSE) = \sigma^2, \quad E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2.$$ 

Since

$$E(MSR) = E \left\{ \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \right\} = E \left\{ \sum_{i=1}^{n} b_1^2(X_i - \bar{X})^2 \right\},$$
and

\[ E(b_1^2) = Var(b_1) + E(b_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \beta_1^2. \]

So if \( \beta_1 = 0 \), we should expect that \( MSE \) and \( MSR \) are about the same magnitude.

The breakdown of variance and degree of freedom are commonly displayed in the form of an analysis of variance table (ANOVA table). We also define the total uncorrected sum of squares, \( SSTOU \), and the correction for the mean sum of squares, \( SS(\text{correction for mean}) \), as

\[ SSTOU = \sum_{i=1}^{n} Y_i^2, \quad SS(\text{correction for mean}) = n\bar{Y}^2. \]
**Analysis of Variance Test**

The two mean square errors suggest the following statistic

\[ F^* = \frac{MSR}{MSE} \]

for testing \( H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0 \).

We have shown that

\[ MSR = b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2. \]

According to (3), when \( \beta_1 = 0 \)

\[ F^* \sim F(1, n - 2), \]
so the rejection region for $\alpha$ Type-I error is constructed as

$$F^* > F(1 - \alpha, 1, n - 2).$$
Degree of Linear Association between $X$ and $Y$

One number to capture the essential information as to whether a given regression relation is useful: coefficient of determination

$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \in [0, 1],$$

which measures the proportionate reduction of total variation associated with the use of the predictor variable $X$.

If $r^2 = 1$, $SSE = 0$. All observations fall on the fitted regression line and $X$ accounts for all variation in the observations $Y_i$.

When $b_1 = 0$, $SSR = 0$ and $r^2 = 0$. There is no linear association between $X$ and $Y$ in the sample data. $X$ is of no help in reducing the
variation in $Y_i$.

The closer $r^2$ is to 1, the greater is said to be the degree of linear association between $X$ and $Y$.

$$r = \text{sign}(b_1) \sqrt{r^2}$$

is called the coefficient of correlation, which is also equal to

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{s_X}{s_Y} b_1, \quad (26)$$

where

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1}}, \quad s_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n - 1}}$$

are the sample standard deviations for the $X$ and $Y$ observations.
Limitations of $r$ and $r^2$

Note that no single measure will be adequate for describing the usefulness of a regression model for different applications.

- $r^2$ measures a *relative reduction* from $SSTO$, and provides no information about absolute precision for estimating a mean response or predicting a new observation.

- $r^2$, $r$ measure the degree of *linear association* between $X$ and $Y$, whereas the actual regression relation may be nonlinear. So high $r$ doesn’t necessarily mean the estimated regression line is a good fit. And $r = 0$ doesn’t necessarily mean $X$ and $Y$ are not related.


Random $X$

In our normal error regression model (1), $X$ are treated as known constants. All the statistical inferences, e.g. estimated confidence coefficients, refer to repeated sampling of $Y$ when keeping $X$ values fixed. When both $X$ and $Y$ are random variables, previous inference results still apply if we assume

1. The conditional distributions of the $Y_i$, given $X_i$, are independent $N(\beta_0 + \beta_1 X, \sigma^2)$.

2. The $X_i$ are independent random variables, and the probability distribution $g(X_i)$ doesn’t involve the regression parameters $\beta_0, \beta_1$ and $\sigma^2$. 