Inference in Regression Analysis

Yang Feng

Inference in the Normal Error Regression Model

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

- Y_i value of the response variable in the i^{th} trial
- β_0 and β_1 are parameters
- X_i is a known constant, the value of the predictor variable in the i^{th} trial
- $\epsilon_i \sim_{iid} N(0, \sigma^2)$
- $i = 1, \ldots, n$

Maximum Likelihood Estimator(s)

- β_0
  b_0 same as in least squares case
- β_1
  b_1 same as in least squares case
- σ^2

$$\hat{\sigma}^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n}$$

- Note that ML estimator is biased as s^2 is unbiased and

$$s^2 = MSE = \frac{n}{n-2}\hat{\sigma}^2$$

Inference concerning β_1

Tests concerning β_1 (the slope) are often of interest, particularly

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0 \end{aligned}$$

the null hypothesis model

$$Y_i = \beta_0 + (0)X_i + \epsilon_i$$

implies that there is no linear relationship between Y and X.

Note the means of all the Y_i's are equal at all levels of X_i.

Sampling Dist. Of b_1

- The point estimator for b_1 is

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

- The sampling distribution for b_1 is the distribution of b_1 that arises from the variability of b_1 when the predictor variables X_i are held fixed and the observed outputs are repeatedly sampled
- Note that the sampling distribution of b_1 will depend on our model assumptions.

Sampling Dist. Of b_1 In Normal Regr. Model

- For a normal error regression model the sampling distribution of b_1 is normal, with mean and variance given by

$$\begin{aligned} \mathbb{E}(b_1) &= \beta_1 \\ \text{Var}(b_1) &= \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \end{aligned}$$

- To show this we need to go through a number of algebraic steps.

First step

To show

$$\sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum(X_i - \bar{X})Y_i$$

we observe

$$\begin{aligned}
\sum(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum(X_i - \bar{X})Y_i - \sum(X_i - \bar{X})\bar{Y} \\
&= \sum(X_i - \bar{X})Y_i - \bar{Y}\sum(X_i - \bar{X}) \\
&= \sum(X_i - \bar{X})Y_i
\end{aligned}$$

b_1 as convex combination of Y_i's

b_1 can be expressed as a linear combination of the $Y_i's$

$$\begin{aligned} b_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \\ &= \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2} \quad \text{from previous slide} \\ &= \sum k_i Y_i \end{aligned}$$

where

$$k_i = \frac{(X_i - \bar{X})}{\sum(X_i - \bar{X})^2}$$

Properties of the $k_i's$

It can be shown that

$$\begin{aligned} \sum k_i &= 0 \\ \sum k_i X_i &= 1 \\ \sum k_i^2 &= \frac{1}{\sum (X_i - \bar{X})^2} \end{aligned}$$

We will use these properties to prove various properties of the sampling distributions of b_1 and b_0.

Linear combination of independent normal random variables

When $Y_1, \ldots, Y_n$ are independent normal random variables, the linear combination

$$\sum a_i Y_i \sim \mathcal{N}\left(\sum a_i \mathbb{E}(Y_i), \sum a_i^2 \operatorname{Var}(Y_i)\right)$$

Normality of $b_1's$ Sampling Distribution

Since b_1 is a linear combination of the $Y_i's$ and each Y_i is an independent normal random variable, then b_1 is distributed normally as well

$$b_1 = \sum k_i Y_i,\ k_i = \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$

From previous slide

$$\mathbb{E}(b_1) = \sum k_i \mathbb{E}(Y_i),\ \ \mathrm{Var}(b_1) = \sum k_i^2 \mathrm{Var}(Y_i)$$

b_1 is an unbiased estimator

This can be seen using two of the properties

$$\begin{aligned}
\mathbb{E}(b_1) &= \mathbb{E}(\sum k_i Y_i) \\
&= \sum k_i \mathbb{E}(Y_i) \\
&= \sum k_i(\beta_0 + \beta_1 X_i) \\
&= \beta_0 \sum k_i + \beta_1 \sum k_i X_i \\
&= \beta_0(0) + \beta_1(1) \\
&= \beta_1
\end{aligned}$$

Variance of b_1

Since the Y_i are independent random variables with variance σ^2 and the $k_i's$ are constants we get

$$\begin{aligned}
\text{Var}(b_1) &= \text{Var}(\sum k_i Y_i) \\
&= \sum k_i^2 \text{Var}(Y_i) \\
&= \sum k_i^2 \sigma^2 \\
&= \sigma^2 \sum k_i^2 \\
&= \sigma^2 \frac{1}{\sum (X_i - \bar{X})^2}
\end{aligned}$$

However, in most cases, σ^2 is unknown.

Estimated variance of b_1

- When we don't know σ^2 then we have to replace it with the MSE estimate (From the Least Square estimation)
- Let

$$s^2 = MSE = \frac{SSE}{n-2}$$

where

$$SSE = \sum e_i^2$$

and

$$e_i = Y_i - \hat{Y}_i$$

plugging in we get

$$\widehat{\text{Var}}(b_1) = \frac{s^2}{\sum (X_i - \bar{X})^2}$$

Recap

- We now have an expression for the sampling distribution of b_1 when σ^2 is known

$$b_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{\sum(X_i - \bar{X})^2}) \tag{1}$$

- When σ^2 is unknown we have an unbiased point estimator of σ^2

$$\widehat{\text{Var}}(b_1) = \frac{s^2}{\sum(X_i - \bar{X})^2}$$

Digression : Gauss-Markov Theorem

Theorem

In a regression model where $\mathbb{E}(\epsilon_i) = 0$ and variance $\text{Var}(\epsilon_i) = \sigma^2 < \infty$ and ϵ_i and ϵ_j are uncorrelated for all i and j the least squares estimators b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators.

No normality assumption on the error distribution!
Recall

$$\begin{aligned} b_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \\ b_0 &= \bar{Y} - b_1\bar{X} \end{aligned}$$

Proof

- The theorem states that b_1 as minimum variance among all unbiased linear estimators of the form

$$\hat{\beta}_1 = \sum c_i Y_i$$

- As this estimator must be unbiased we have

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \sum c_i \, \mathbb{E}(Y_i) \\
&= \sum c_i (\beta_0 + \beta_1 X_i) \\
&= \beta_0 \sum c_i + \beta_1 \sum c_i X_i \\
&= \beta_1
\end{aligned}$$

Proof cont.

- Given the constraint

$$\beta_0 \sum c_i + \beta_1 \sum c_i X_i = \beta_1$$

  clearly it must be the case that $\sum c_i = 0$ and $\sum c_i X_i = 1$
- The variance of this estimator is

$$\text{Var}(\hat{\beta}_1) = \sum c_i^2 \text{Var}(Y_i) = \sigma^2 \sum c_i^2$$

Proof cont.

Now define $c_i = k_i + d_i$ where the k_i are the constants we already defined and the d_i are arbitrary constants.

$$k_i = \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$

Let's look at the variance of the estimator

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \sum c_i^2 \, \text{Var}(Y_i) = \sigma^2 \sum (k_i + d_i)^2 \\ &= \sigma^2 (\sum k_i^2 + \sum d_i^2 + 2 \sum k_i d_i) \end{aligned}$$

Note we just demonstrated that

$$\sigma^2 \sum k_i^2 = \text{Var}(b_1)$$

Proof cont.

Now by showing that $\sum k_i d_i = 0$ we're almost done

$$\begin{aligned}
\sum k_i d_i &= \sum k_i (c_i - k_i) \\
&= \sum k_i c_i - \sum k_i^2 \\
&= \sum c_i \left(\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right) - \frac{1}{\sum (X_i - \bar{X})^2} \\
&= \frac{\sum c_i X_i - \bar{X} \sum c_i}{\sum (X_i - \bar{X})^2} - \frac{1}{\sum (X_i - \bar{X})^2} = 0
\end{aligned}$$

Recall $\sum c_i = 0$ and $\sum c_i X_i = 1$.

Proof end

So we are left with

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \sigma^2(\sum k_i^2 + \sum d_i^2) \\ &= \text{Var}(b_1) + \sigma^2(\sum d_i^2) \end{aligned}$$

which is minimized when all the $d_i = 0$. This means that the least squares estimator b_1 has minimum variance among all unbiased linear estimators.

Sampling Distribution of $(b_1 - \beta_1)/s(b_1)$

- b_1 is normally distributed so

$$\frac{b_1 - \beta_1}{\sqrt{\text{Var}(b_1)}} \sim N(0, 1)$$

- We don't know $\text{Var}(b_1)$ so it must be estimated from data. We have already derived its estimate
- If using the estimate $\widehat{\text{Var}}(b_1)$ it can be shown that

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

$$s(b_1) = \sqrt{\widehat{\text{Var}}(b_1)}$$

Where does this come from?

- For now we need to rely upon the following theorem

  For the normal error regression model

$$\frac{SSE}{\sigma^2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2(n-2)$$

  and is independent of b_0 and b_1

- Intuitively this follows the standard result for the sum of squared normal random variables
- Here there are two linear constraints imposed by the regression parameter estimation that each reduce the number of degrees of freedom by one.
- We will revisit this subject soon.

Another useful fact : t distributed random variables

Let z and $\chi^2(\nu)$ be independent random variables (standard normal and χ^2 respectively). The following random variable is a t-dstributed random variable:

$$t(\nu) = \frac{z}{\sqrt{\frac{\chi^2(\nu)}{\nu}}}$$

This version of the t distribution has one parameter, the degrees of freedom ν

Distribution of the studentized statistic

To derive the distribution of this statistic, first we rewrite

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

$$\frac{s(b_1)}{\sigma(b_1)} = \sqrt{\frac{\widehat{\text{Var}}(b_1)}{\text{Var}(b_1)}}$$

Studentized statistic cont.

And note the following

$$\frac{\widehat{\text{Var}}(b_1)}{\text{Var}(b_1)} = \frac{\frac{MSE}{\sum(X_i-\bar{X})^2}}{\frac{\sigma^2}{\sum(X_i-\bar{X})^2}} = \frac{MSE}{\sigma^2} = \frac{SSE}{\sigma^2(n-2)}$$

where we know (by the given theorem) the distribution of the last term is χ^2 and indep. of b_1 and b_0

$$\frac{SSE}{\sigma^2(n-2)} \sim \frac{\chi^2(n-2)}{n-2}$$

Studentized statistic final

But by the given definition of the t distribution we have our result

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

because putting everything together we can see that

$$\frac{b_1 - \beta_1}{s(b_1)} \sim \frac{z}{\sqrt{\frac{\chi^2(n-2)}{n-2}}}$$

Confidence Intervals and Hypothesis Tests

Now that we know the sampling distribution of b_1 (t with $n-2$ degrees of freedom) we can construct confidence intervals and hypothesis tests easily.

Things to think about

- What does the t-distribution look like?
- Why is the estimator distributed according to a t-distribution rather than a normal distribution?
- When performing tests, why does this matter?
- When is it safe to cheat and use a normal approximation?

Quick Review : Hypothesis Testing

- Elements of a statistical test
  - Null hypothesis, H_0
  - Alternative hypothesis, H_a
  - Test statistic
  - Rejection region

Quick Review : Hypothesis Testing - Errors

- Errors
  - A type I error is made if H_0 is rejected when H_0 is true. The probability of a type I error is denoted by α. The value of α is called the level of the test.
  - A type II error is made if H_0 is accepted when H_a is true. The probability of a type II error is denoted by β.

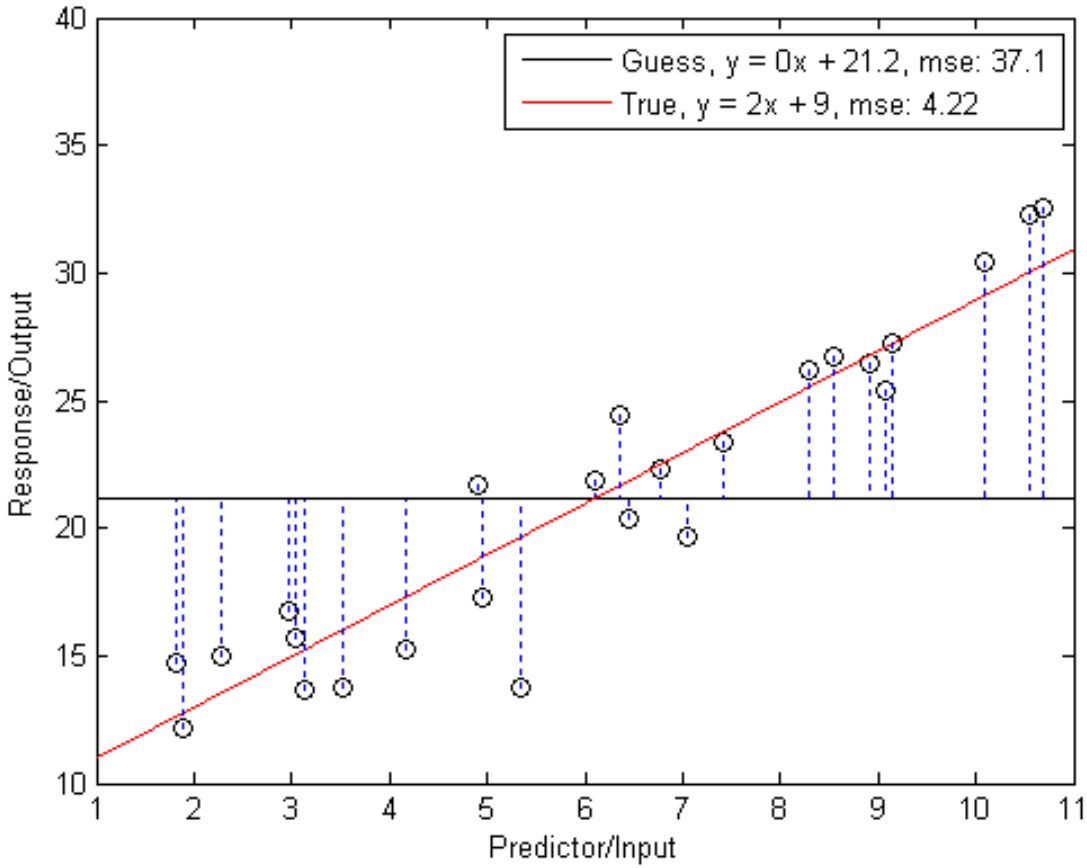
	H_0 is true	H_a is true
Accept H_0	Right decision	Type II Error
Reject H_0	Type I Error	Right decision

p-value

The p-value, or attained significance level, is the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected.

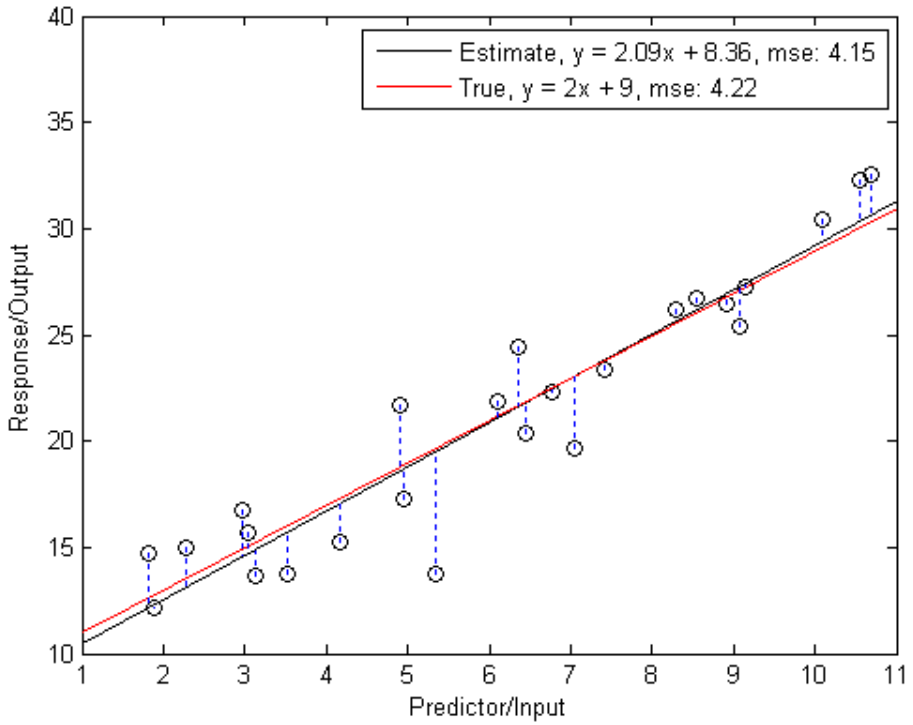
Null Hypothesis

If the null hypothesis is that $\beta_1 = 0$ then b_1 should fall in the range around zero. The further it is from 0 the less likely the null hypothesis is to hold.



Alternative Hypothesis : Least Squares Fit

If we find that our estimated value of b_1 deviates from 0 then we have to determine whether or not that deviation would be surprising given the model and the sampling distribution of the estimator. If it is sufficiently (where we define what sufficient is by a confidence level) different then we reject the null hypothesis.

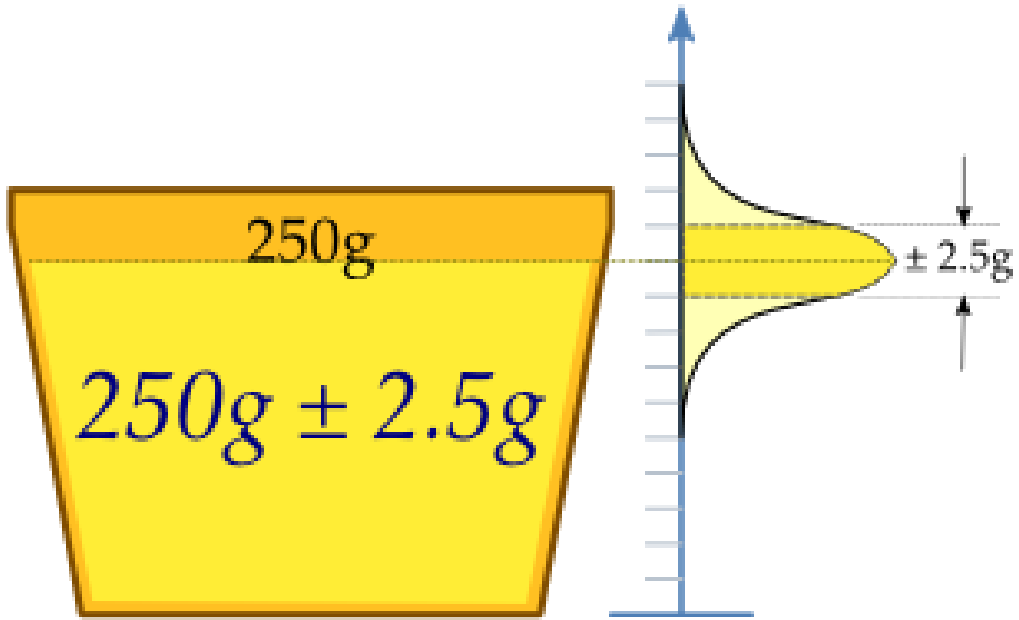


Testing This Hypothesis

- Only have a finite sample
- Different finite set of samples (from the same population / source) will (almost always) produce different point estimates of β_0 and β_1 (b_0, b_1) given the same estimation procedure
- Key point: b_0 and b_1 are random variables whose sampling distributions can be statistically characterized
- Hypothesis tests about β_0 and β_1 can be constructed using these distributions.

Confidence Interval Example

- A machine fills cups with margarine, and is supposed to be adjusted so that the content of the cups is $\mu = 250g$ of margarine.
- Observed random variable $X \sim \mathcal{N}(250, 2.5)$



Confidence Interval Example, Cont.

- $X_1, ..., X_{25}$, a random sample from X.
- The natural estimator is the sample mean: $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- Suppose the sample shows actual weights $X_1, ..., X_{25}$, with mean:

$$\bar{X} = \frac{1}{25}\sum_{i=1}^{25} X_i = 250.2\text{grams}.$$

Confidence Interval Example, Cont.

Say we want to get a confidence interval for μ.
By standardizing, we get a random variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{0.5}$$

$$P(-z \leq Z \leq z) = 1 - \alpha = 0.95.$$

The number z follows from the cumulative distribution function:

$$\Phi(z) = P(Z \leq z) = 1 - \tfrac{\alpha}{2} = 0.975, \tag{2}$$

$$z = \Phi^{-1}(\Phi(z)) = \Phi^{-1}(0.975) = 1.96, \tag{3}$$

Confidence Interval Example, Cont.

Now we get:

$$0.95 = 1 - \alpha = P(-z \leq Z \leq z) = P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \quad (4)$$

$$= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \quad (5)$$

$$= P\left(\bar{X} - 1.96 \times 0.5 \leq \mu \leq \bar{X} + 1.96 \times 0.5\right) \quad (6)$$

$$= P\left(\bar{X} - 0.98 \leq \mu \leq \bar{X} + 0.98\right). \quad (7)$$

This might be interpreted as: with probability 0.95 we will find a confidence interval in which we will meet the parameter μ between the stochastic endpoints

$$(\bar{X} - 0.98, \bar{X} + 0.98)$$

Confidence Interval Example, Cont.

Therefore, our 0.95 confidence interval becomes:

$$(\bar{X} - 0.98, \bar{X} + 0.98) = (250.2 - 0.98, 250.2 + 0.98) = (249.22, 251.18).$$

Recap

- We know that the point estimator of β_1 is

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

- We derived the sampling distribution of b_1, it being $\mathcal{N}(\beta_1, Var(b_1))$(when σ^2 known) with

$$\text{Var}(b_1) = \sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

- And we suggested that an estimate of $\text{Var}(b_1)$ could be arrived at by substituting the MSE for σ^2 when σ^2 is unknown.

$$s^2\{b_1\} = \frac{MSE}{\sum(X_i - \bar{X})^2} = \frac{\frac{SSE}{n-2}}{\sum(X_i - \bar{X})^2}$$

Sampling Distribution of $(b_1 - \beta_1)/s\{b_1\}$

- Since b_1 is normally distributed,

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

- Using the estimate $s^2\{b_1\}$,

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

where

$$s\{b_1\} = \sqrt{s^2\{b_1\}}$$

Confidence Interval for β_1

Since the "studentized" statistic follows a t distribution, we can make the following probability statement

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

By symmetry

$$t(\alpha/2; n-2) = -t(1-\alpha/2; n-2),$$

We have the following confidence interval for β_1,

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

Look up the t distribution table (table B.2 in the appendix) and produce confidence intervals.
Or R function

$$qt(1-\alpha/2, n-2)$$

Remember

- Density: $f(y) = \frac{dF(y)}{dy}$
- Distribution (CDF): $F(y) = P(Y \leq y) = \int_{-\infty}^{y} f(t)dt$
- Inverse CDF: $F^{-1}(p) = y$ s.t. $\int_{-\infty}^{y} f(t)dt = p$

$1-\alpha$ confidence limits for β_1

- The $1-\alpha$ confidence limits for β_1 are

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

- Note that this quantity can be used to calculate confidence intervals given n and α.
  - Fixing α can guide the choice of sample size if a particular confidence interval is desired
  - Given a sample size, vice versa.
- Also useful for hypothesis testing

Tests Concerning β_1

- Example 1(Two-sided test)
  - $H_0 : \beta_1 = 0$
  - $H_a : \beta_1 \neq 0$
  - Test statistic

$$t^* = \frac{b_1 - 0}{s\{b_1\}}$$

Tests Concerning β_1

- We have an estimate of the sampling distribution of b_1 from the data.
- If the null hypothesis holds then the b_1 estimate coming from the data should be within the 95% confidence interval of the sampling distribution centered at 0 (in this case)

$$t^* = \frac{b_1 - 0}{s\{b_1\}}$$

Decision rules

$$\begin{aligned} \text{if } |t^*| &\leq t(1-\alpha/2; n-2), \text{accept } H_0 \\ \text{if } |t^*| &> t(1-\alpha/2; n-2), \text{reject } H_0 \end{aligned}$$

Absolute values make the test two-sided

Calculating the p-value

- The p-value, or attained significance level, is the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected.
- This can be looked up using the CDF of the test statistic.

p-value Example

An experiment is performed to determine whether a coin flip is fair (50% chance, each, of landing heads or tails) or unfairly biased (50% chance of one of the outcomes).

Outcome: Suppose that the experimental results show the coin turning up heads 14 times out of 20 total flips.
p-value: The p-value of this result would be the chance of a fair coin landing on heads at least 14 times out of 20 flips
Calculation:

$$\text{Prob}(14 \text{ heads}) + \text{Prob}(15 \text{ heads}) + \cdots + \text{Prob}(20 \text{ heads}) \tag{8}$$

$$= \frac{1}{2^{20}} \left[\binom{20}{14} + \binom{20}{15} + \cdots + \binom{20}{20} \right] = \frac{60{,}460}{1{,}048{,}576} \approx 0.058 \tag{9}$$

Two sided p-value:

$$2 * 0.058 = 0.116$$

Power of the test

$$\text{Power} = P\{\text{Reject } H_0|\delta\} \tag{10}$$
$$= P\{|t^*| > t(1-\alpha/2; n-2)|\delta\} \tag{11}$$
$$\tag{12}$$

where δ is the *noncentrality measure*

$$\delta = \frac{|\beta_1|}{\sigma\{b_1\}}$$

$$t^* = \frac{\frac{b_1-\beta_1}{\sigma(b_1)} + \frac{\beta_1}{\sigma(b_1)}}{\frac{s(b_1)}{\sigma(b_1)}} \tag{13}$$

Table B.5 presents the power of the two-sided t test.
Notice: the power depends on the value of σ^2.

```
a = qt(1- alpha/2,n-2)
1- pt(a, n-2, delta) + pt(-a, n-2, delta)
```

Inferences Concerning β_0

- Largely, inference procedures regarding β_0 can be performed in the same way as those for β_1
- Remember the point estimator b_0 for β_0

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

Sampling distribution of b_0

- The sampling distribution of b_0 refers to the different values of b_0 that would be obtained with repeated sampling when the levels of the predictor variable X are held constant from sample to sample.
- For the normal regression model the sampling distribution of b_0 is normal

Sampling distribution of b_0

- When error variance is known

$$\mathbb{E}(b_0) = \beta_0$$

$$\sigma^2\{b_0\} = \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2})$$

- When error variance is unknown

$$s^2\{b_0\} = MSE(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2})$$

Confidence interval for β_0

The $1-\alpha$ confidence limits for β_0 are obtained in the same manner as those for β_1

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

Considerations on Inferences on β_0 and β_1

- Effects of departures from normality of the Y_i
  The estimators of β_0 and β_1 have the property of asymptotic normality - their distributions approach normality as the sample size increases (under general conditions)
- Spacing of the X levels
  The variances of b_0 and b_1 (for a given n and σ^2) depend strongly on the spacing of X

Sampling distribution of point estimator of mean response

- Let X_h be the level of X for which we would like an estimate of the mean response
  Needs to be one of the observed X's
- The mean response when $X = X_h$ is denoted by

$$\mathbb{E}(Y_h) = \beta_0 + \beta_1 X_h$$

- The point estimator of $\mathbb{E}(Y_h)$ is

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

We are interested in the sampling distribution of this quantity

Sampling Distribution of $\hat{Y}_h$

- We have
$$\hat{Y}_h = b_0 + b_1 X_h$$
- Since this quantity is itself a linear combination of the $Y_i's$ it's sampling distribution is itself normal.
- The mean of the sampling distribution is
$$E\{\hat{Y}_h\} = E\{b_0\} + E\{b_1\}X_h = \beta_0 + \beta_1 X_h$$

Biased or unbiased?

Sampling Distribution of $\hat{Y}_h$

- To derive the sampling distribution variance of the mean response we first show that b_1 and $(1/n)\sum Y_i$ are uncorrelated and, hence, for the normal error regression model independent
- We start with the definitions

$$\bar{Y} = \sum (\frac{1}{n}) Y_i$$

$$b_1 = \sum k_i Y_i, \; k_i = \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$

Sampling Distribution of $\hat{Y}_h$

- We want to show that mean response and the estimate b_1 are uncorrelated

$$Cov(\bar{Y}, b_1) = \sigma^2\{\bar{Y}, b_1\} = 0$$

- To do this we need the following result (A.32)

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

when the Y_i are independent

Sampling Distribution of $\hat{Y}_h$

Using this fact we have

$$\begin{aligned}
\sigma^2\{\sum_{i=1}^{n}\frac{1}{n}Y_i, \sum_{i=1}^{n}k_iY_i\} &= \sum_{i=1}^{n}\frac{1}{n}k_i\sigma^2\{Y_i\} \\
&= \sum_{i=1}^{n}\frac{1}{n}k_i\sigma^2 \\
&= \frac{\sigma^2}{n}\sum_{i=1}^{n}k_i \\
&= 0
\end{aligned}$$

So the $\bar{Y}$ and b_1 are uncorrelated

Sampling Distribution of $\hat{Y}_h$

- This means that we can write down the variance

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\{\bar{Y} + b_1(X_h - \bar{X})\}$$

  alternative and equivalent form of regression function
- But we know that $\bar{Y}$ and b_1 are uncorrelated so

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\{\bar{Y}\} + \sigma^2\{b_1\}(X_h - \bar{X})^2$$

Sampling Distribution of $\hat{Y}_h$

- We know

$$\begin{aligned} \sigma^2\{b_1\} &= \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \\ s^2\{b_1\} &= \frac{MSE}{\sum(X_i - \bar{X})^2} \end{aligned}$$

- And we can find

$$\sigma^2\{\bar{Y}\} = \frac{1}{n^2}\sum \sigma^2\{Y_i\} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Sampling Distribution of $\hat{Y}_h$

- So, plugging in, we get

$$\sigma^2\{\hat{Y}_h\} = \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum(X_i - \bar{X})^2}(X_h - \bar{X})^2$$

- Or

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)$$

Sampling Distribution of $\hat{Y}_h$

Since we often won't know σ^2 we can, as usual, plug in $s^2 = SSE/(n-2)$, our estimate for it to get our estimate of this sampling distribution variance

$$s^2\{\hat{Y}_h\} = s^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)$$

No surprise...

- The sampling distribution of our point estimator for the output is distributed as a t-distribution with $n-2$ degrees of freedom

$$\frac{\hat{Y}_h - E\{Y_h\}}{s\{\hat{Y}_h\}} \sim t(n-2)$$

- This means that we can construct confidence intervals in the same manner as before.

Confidence Intervals for $\mathbb{E}(Y_h)$

- The $1-\alpha$ confidence intervals for $\mathbb{E}(Y_h)$ are

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

- From this hypothesis tests can be constructed as usual.

Comments

- The variance of the estimator for $\mathbb{E}(Y_h)$ is smallest near the mean of X. Designing studies such that the mean of X is near X_h will improve inference precision
- When X_h is zero the variance of the estimator for $\mathbb{E}(Y_h)$ reduces to the variance of the estimator b_0 for β_0

Confidence Band for Regression Line

- At times, we want to get a confidence band for the entire regression line $E\{Y\} = \beta_0 + \beta_1 X$.
- The Working-Hotelling $1 - \alpha$ confidence band is

$$\hat{Y}_h \pm W \times s\{\hat{Y}_h\}$$

  here $W^2 = 2F(1 - \alpha; 2, n - 2)$.
- Same form as before, except the t multiple is replaced with the W multiple.

R code:

$$qf(1 - \alpha/2, 2, n - 2)$$

Prediction interval for single new observation

- Essentially follows the sampling distribution arguments for $\mathbb{E}(Y_h)$
- If all regression parameters are known then the $1-\alpha$ prediction interval for a new observation Y_h is

$$\mathbb{E}\{Y_h\} \pm z(1-\alpha/2)\sigma$$

Prediction interval for single new observation

- If the regression parameters are unknown the $1-\alpha$ prediction interval for a new observation Y_h is given by the following theorem

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

- This is very nearly the same as prediction for a known value of X but includes a correction for the fact that there is additional variability arising from the fact that the new input location was not used in the original estimates of b_1, b_0, and s^2

Prediction interval for single new observation

We have

$$\sigma^2\{pred\} = \sigma^2\{Y_h - \hat{Y}_h\} = \sigma^2\{Y_h\} + \sigma^2\{\hat{Y}_h\} = \sigma^2 + \sigma^2\{\hat{Y}_h\}$$

An unbiased estimator of $\sigma^2\{pred\}$ is $s^2\{pred\} = MSE + s^2\{\hat{Y}_h\}$, which is given by

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

- Confidence Interval: represents an inference on a parameter, and is an interval which is intended to cover the value of the parameter.
- Prediction Interval: a statement about the value to be taken by a random variable. Wider than confidence interval.

Prediction interval for the mean of m new observations for given X_h

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

An unbiased estimator of $\sigma^2\{pred\}$ is $s^2\{predmean\} = \frac{MSE}{m} + s^2\{\hat{Y}_h\}$, which is given by

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

ANOVA

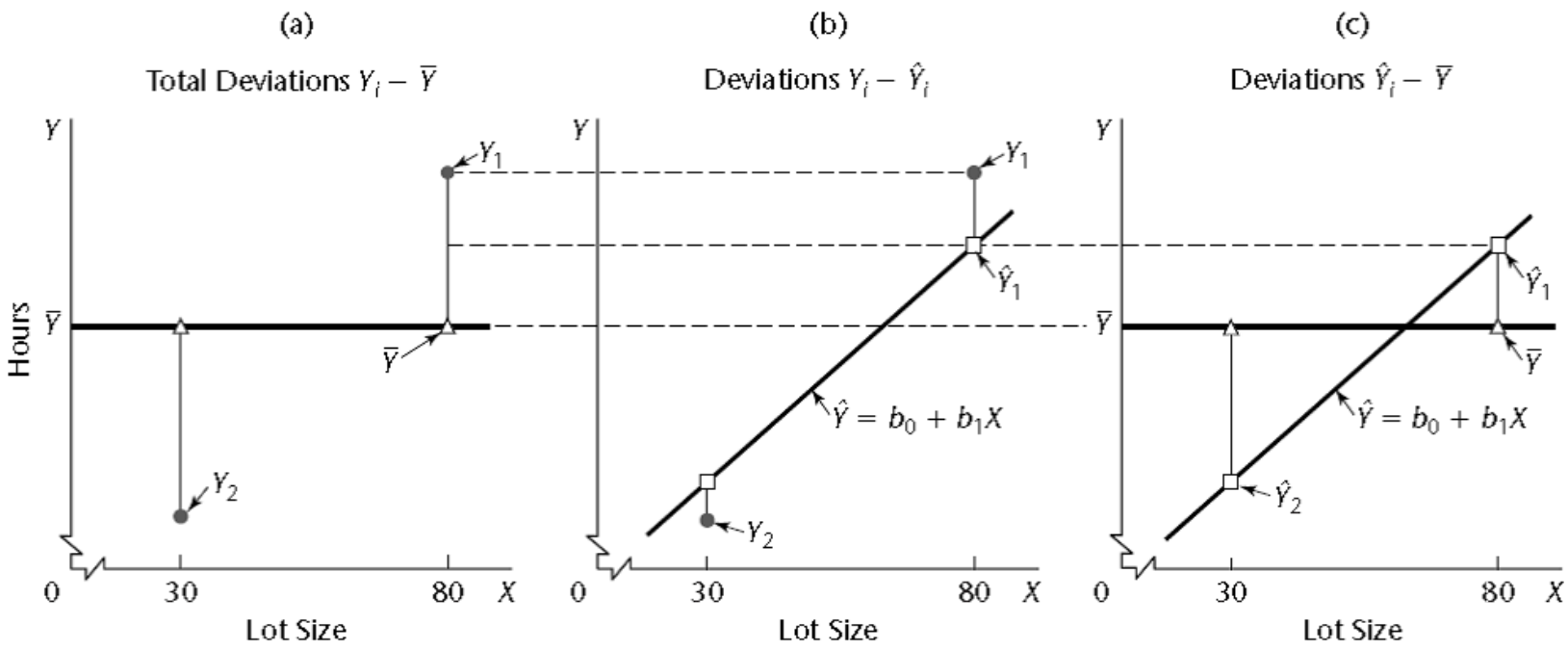
- ANOVA is nothing new but is instead a way of organizing the parts of linear regression so as to make easy inference recipes.
- Will return to ANOVA when discussing multiple regression and other types of linear statistical models.

Partitioning Total Sum of Squares

- "The ANOVA approach is based on the partitioning of sums of squares and degrees of freedom associated with the response variable Y"
- We start with the observed deviations of Y_i around the observed mean

$$Y_i - \bar{Y}$$

Partitioning of Total Deviations



Measure of Total Variation

- The measure of total variation is denoted by

$$SSTO = \sum (Y_i - \bar{Y})^2$$

- SSTO stands for total sum of squares
- If all $Y_i's$ are the same, SSTO $= 0$
- The greater the variation of the $Y_i's$ the greater SSTO

Variation after predictor effect

- The measure of variation of the $Y_i's$ that is still present when the predictor variable X is taken into account is the sum of the squared deviations

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

- SSE denotes error sum of squares

Regression Sum of Squares

- The difference between SSTO and SSE is SSR

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2$$

- SSR stands for regression sum of squares

Partitioning of Sum of Squares

$$\underbrace{Y_i - \bar{Y}}_{\text{Total deviation}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{Deviation of fitted regression value around mean}} + \underbrace{Y_i - \hat{Y}_i}_{\text{Deviation around fitted regression line}}$$

Remarkable Property

- The sums of the same deviations squared has the same property!

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

or

$$SSTO = SSR + SSE$$

Proof

$$\begin{aligned}\sum(Y_i-\bar{Y})^2 &= \sum[(\hat{Y}_i-\bar{Y})+(Y_i-\hat{Y}_i)]^2 \\ &= \sum[(\hat{Y}_i-\bar{Y})^2+(Y_i-\hat{Y}_i)^2+2(\hat{Y}_i-\bar{Y})(Y_i-\hat{Y}_i)] \\ &= \sum(\hat{Y}_i-\bar{Y})^2+\sum(Y_i-\hat{Y}_i)^2+2\sum(\hat{Y}_i-\bar{Y})(Y_i-\hat{Y}_i)\end{aligned}$$

but

$$\sum(\hat{Y}_i-\bar{Y})(Y_i-\hat{Y}_i)=\sum\hat{Y}_i(Y_i-\hat{Y}_i)-\sum\bar{Y}(Y_i-\hat{Y}_i)=0$$

By properties previously demonstrated. Namely

$$\sum\hat{Y}_i e_i=0$$

and

$$\sum e_i=0$$

Breakdown of Degrees of Freedom

- SSTO
  - 1 linear constraint due to the calculation and inclusion of the mean
    - n-1 degrees of freedom
- SSE
  - 2 linear constraints arising from the estimation of β_1 and β_0
    - n-2 degrees of freedom
- SSR
  - Two degrees of freedom in the regression parameters, one is lost due to linear constraint
    - 1 degree of freedom

Mean Squares

A sum of squares divided by its associated degrees of freedom is called a mean square

The regression mean square is

$$MSR = \frac{SSR}{1} = SSR$$

The mean square error is

$$MSE = \frac{SSE}{n-2}$$

ANOVA table for simple lin. regression

Source of Variation	SS	df	MS	$\mathbb{E}(MS)$
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = SSR/1$	$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = SSE/(n-2)$	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		

$\mathbb{E}\{MSE\} = \sigma^2$

- Remember the following theorem, presented in an earlier lecture.

  For the normal error regression model, $\frac{SSE}{\sigma^2}$ is distributed as χ^2 with $n-2$ degrees of freedom and is independent of both b_0 and b_1.

  Rewritten this yields

$$SSE/\sigma^2 \sim \chi^2(n-2)$$

- That means that $\mathbb{E}\{SSE/\sigma^2\} = n-2$
- And thus that $\mathbb{E}\{SSE/(n-2)\} = \mathbb{E}\{MSE\} = \sigma^2$

$E\{MSR\} = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$

- To begin, we take an alternative but equivalent form for SSR

$$SSR = b_1^2 \sum (X_i - \bar{X})^2$$

- And note that, by definition of variance we can write

$$\sigma^2\{b_1\} = E\{b_1^2\} - (E\{b_1\})^2$$

$E\{MSR\} = \sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$

- But we know that b_1 is an unbiased estimator of β_1 so $\mathbb{E}\{b_1\} = \beta_1$
- We also know (from previous lectures) that

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

- So we can rearrange terms and plug in

$$\begin{aligned} \sigma^2\{b_1\} &= E\{b_1^2\} - (E\{b_1\})^2 \\ E\{b_1^2\} &= \frac{\sigma^2}{\sum(X_i - \bar{X})^2} + \beta_1^2 \end{aligned}$$

$E\{MSR\} = \sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$

- From the previous slide

$$E\{b_1^2\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} + \beta_1^2$$

- Which brings us to this result
  $E\{MSR\} = E\{SSR/1\} = E\{b_1^2\} \sum(X_i - \bar{X})^2 = \sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$

Comments and Intuition

- The mean of the sampling distribution of MSE is σ^2 regardless of whether X and Y are linearly related (i.e. whether $\beta_1 = 0$)
- The mean of the sampling distribution of MSR is also σ^2 when $\beta_1 = 0$.
  - When $\beta_1 = 0$ the sampling distributions of MSR and MSE tend to be the same

F Test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$

ANOVA provides a battery of useful tests. For example, ANOVA provides an easy test for
Two-sided test

$$H_0 : \beta_1 = 0 \text{ v.s. } H_a : \beta_1 \neq 0$$

Test statistic from before

$$t^* = \frac{b_1 - 0}{s\{b_1\}}$$

ANOVA test statistic

$$F^* = \frac{MSR}{MSE}$$

Sampling distribution of F^*

- The sampling distribution of F^* when $H_0 : \beta_1 = 0$ holds can be derived from Cochran's theorem
- Cochran's theorem
  If all n observations Y_i come from the same normal distribution with mean μ and variance σ^2, and SSTO is decomposed into k sums of squares SS_r, each with degrees of freedom df_r, then the SS_r/σ^2 terms are independent χ^2 variables with df_r degrees of freedom if

$$\sum_{r=1}^{k} df_r = n - 1$$

The F Test

We have decomposed SSTO into two sums of squares SSR and SSE and their degrees of freedom are additive, hence, by Cochran's theorem:
If $\beta_1 = 0$ so that all Y_i have the same mean $\mu = \beta_0$ and the same variance σ^2, SSE/σ^2 and SSR/σ^2 are independent χ^2 variables

F^* Test Statistic

- F^* can be written as follows

$$F^* = \frac{MSR}{MSE} = \frac{\frac{SSR/\sigma^2}{1}}{\frac{SSE/\sigma^2}{n-2}}$$

- But by Cochran' s theorem, we have when H_0 holds

$$F^* \sim \frac{\frac{\chi^2(1)}{1}}{\frac{\chi^2(n-2)}{n-2}}$$

F Distribution

- The F distribution is the ratio of two independent χ^2 random variables normalized by their corresponding degrees of freedom.
- The test statistic F^* follows the distribution

$$F^* \sim F(1, n-2)$$

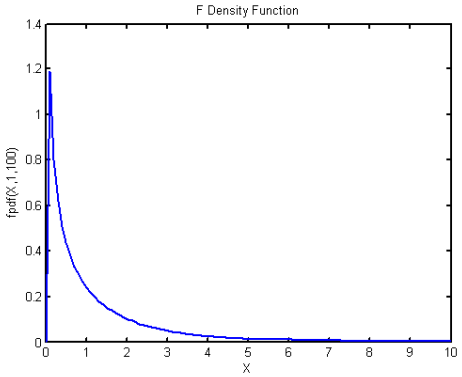
Hypothesis Test Decision Rule

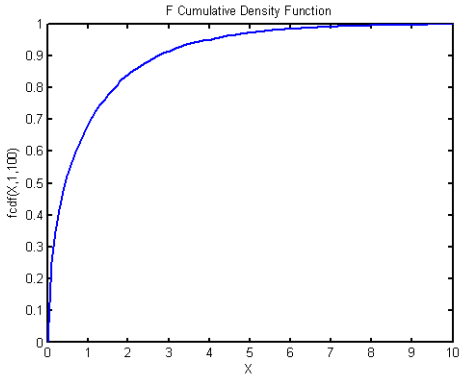
Since F^* is distributed as $F(1, n-2)$ when H_0 holds, the decision rule to follow when the risk of a Type I error is to be controlled at α is:

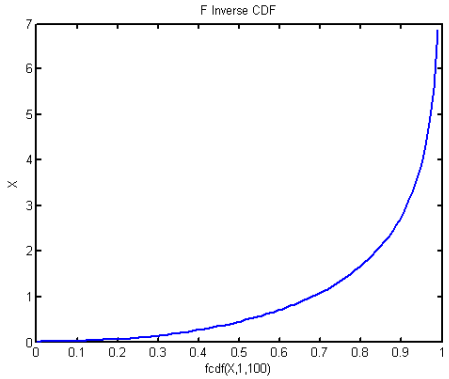
If $F^* \leq F(1-\alpha; 1, n-2)$, conclude H_0
If $F^* > F(1-\alpha; 1, n-2)$, conclude H_a

F distribution

- PDF, CDF, Inverse CDF of F distribution



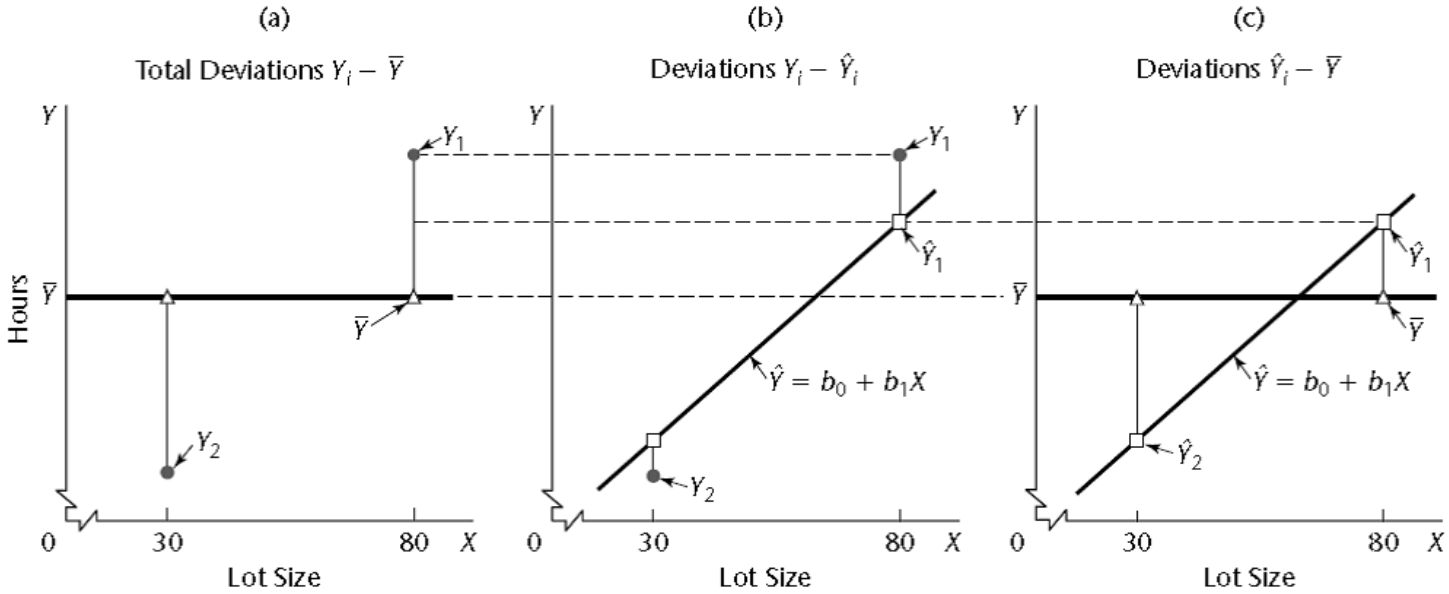




- Note, MSR/MSE must be big in order to reject hypothesis.

Partitioning of Total Deviations

Does this make sense? When is MSR/MSE big?



Equivalence of F test and two-sided t test

$$F^* = \frac{MSR}{MSE} \tag{14}$$

$$= \frac{b_1^2 \sum (X_i - \bar{X})^2}{MSE} \tag{15}$$

$$= \frac{b_1^2}{s^2\{b_1\}} \tag{16}$$

$$= \left(\frac{b_1}{s\{b_1\}}\right)^2 \tag{17}$$

$$= (t^*)^2 \tag{18}$$

In addition: $F(1-\alpha; 1, n-2) = t(1-\alpha/2; n-2)^2$.

General Linear Test

- The test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is a simple example of a general linear test.
- The general linear test has three parts
  - Full Model
  - Reduced Model
  - Test Statistic

Full Model Fit

- A full linear model is first fit to the data

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

- Using this model the error sum of squares is obtained, here for example the simple linear model with non-zero slope is the "full" model

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

Fit Reduced Model

- One can test the hypothesis that a simpler model is a "better" model via a general linear test (which is really a likelihood ratio test in disguise). For instance, consider a "reduced" model in which the slope is zero (i.e. no relationship between input and output).

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

- The model when H_0 holds is called the reduced or restricted model.

$$Y_i = \beta_0 + \epsilon_i$$

- The SSE for the reduced model is obtained

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

Test Statistic

- The idea is to compare the two error sums of squares SSE(F) and SSE(R).
- Because the full model F has more parameters than the reduced model, $SSE(F) \leq SSE(R)$ always
- In the general linear test, the test statistic is

$$F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R-df_F}}{\frac{SSE(F)}{df_F}}$$

  which follows the F distribution when H_0 holds.
- df_R and df_F are those associated with the reduced and full model error sums of squares respectively

R^2 (Coefficient of determination)

- SSTO measures the variation in the observations Y_i when X is not considered
- SSE measures the variation in the Y_i after a predictor variable X is employed
- A natural measure of the effect of X in reducing variation in Y is to express the reduction in variation ($SSTO - SSE = SSR$) as a proportion of the total variation

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- Note that since $0 \leq SSE \leq SSTO$ then $0 \leq R^2 \leq 1$

Limitations of and misunderstandings about R^2

1. Claim: high R^2 indicates that useful predictions can be made. The prediction interval for a particular input of interest may still be wide even if R^2 is high.
2. Claim: high R^2 means that there is a good linear fit between predictor and output. It can be the case that an approximate (bad) linear fit to a truly curvilinear relationship might result in a high R^2.
3. Claim: low R^2 means that there is no relationship between input and output. Also not true since there can be clear and strong relationships between input and output that are not well explained by a linear functional relationship.

Coefficient of Correlation

$$r = \pm\sqrt{R^2}$$

Range:

$$-1 \leq r \leq 1$$

if $b_1 > 0$, $r = \sqrt{R^2}$,
if $b_1 < 0$, $r = -\sqrt{R^2}$.

Normal Correlation Models

Some times, correlation models are more natural than regression models, such as

- Relationship between sales of gasoline and sales of auxiliary products.
- Relationship between blood pressure and age.

Difference between Regression models and Correlation models

- Regression models: X values are known constants.
- Correlation models: Both X and Y are random.

Bivariate Normal Density

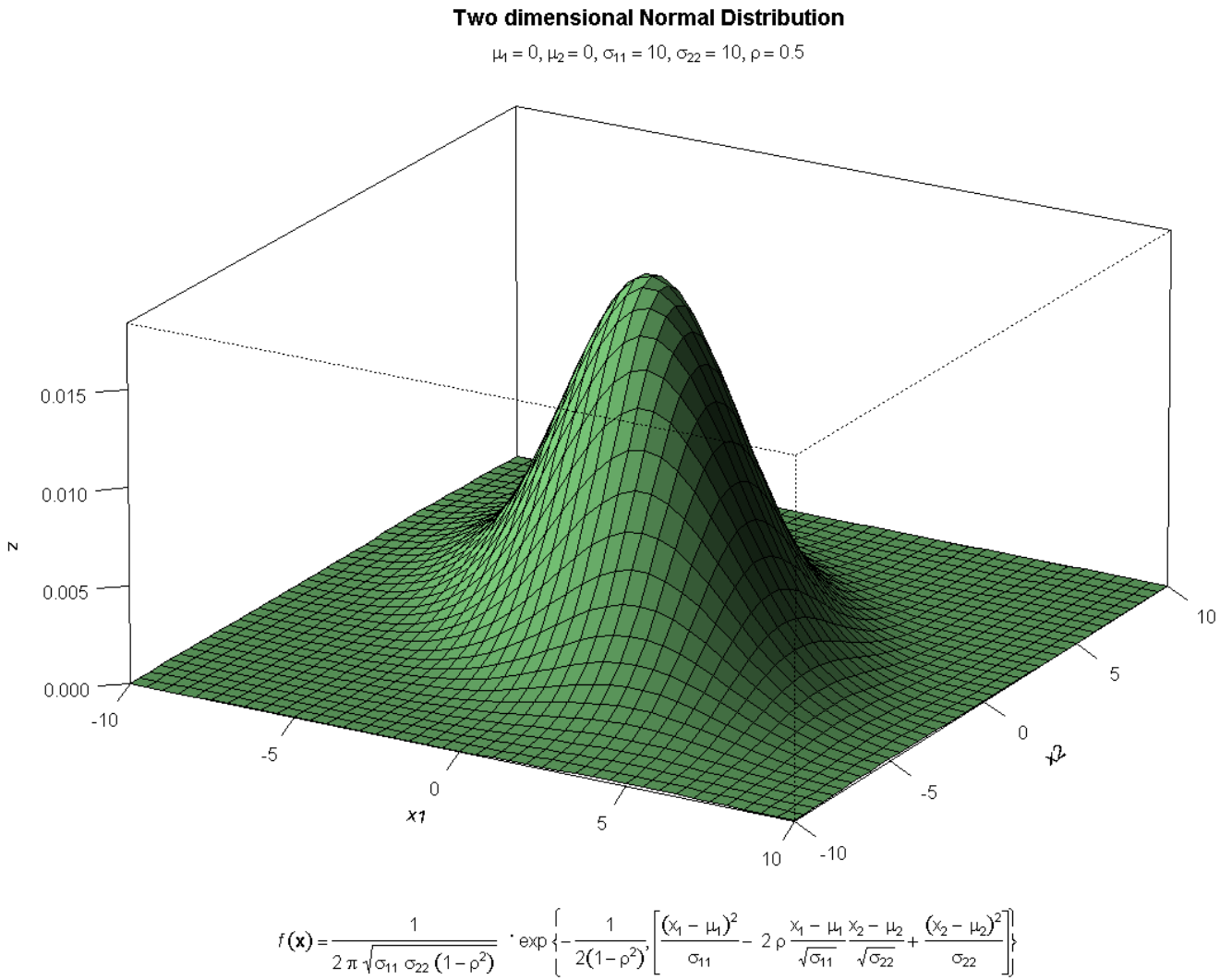
$$f(Y_1, Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} \exp\{-\frac{1}{2(1-\rho_{12}^2)}[(\frac{Y_1-\mu_1}{\sigma_1})^2 - 2\rho_{12}(\frac{Y_1-\mu_1}{\sigma_1})(\frac{Y_2-\mu_2}{\sigma_2}) + (\frac{Y_2-\mu_2}{\sigma_2})^2]\}$$

Parameters: $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12}$

Here, ρ_{12} is the coefficient of correlation between variable Y_1 and Y_2.

$$\rho_{12} = \rho\{Y_1, Y_2\} = \frac{\sigma_{12}}{\sigma_1\sigma_2}$$

$$\sigma_{12} = \sigma\{Y_1, Y_2\} = E\{(Y_1-\mu_1)(Y_2-\mu_2)\}$$



$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{\sigma_{11}\,\sigma_{22}\,(1-\rho^2)}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_{11}} - 2\rho\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} + \frac{(x_2-\mu_2)^2}{\sigma_{22}}\right]\right\}$$

Marginal Density

Marginal distribution of Y_1 is normal with mean μ_1 and standard deviation σ_1:

$$f_1(Y_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp[-\frac{1}{2}(\frac{Y_1 - \mu_1}{\sigma_1})^2]$$

How to get it from the bivariate density function?

Conditional Inferences

$$
\begin{aligned}
f(Y_1|Y_2) &= \frac{f(Y_1, Y_2)}{f_2(Y_2)} \\
&= \frac{1}{\sqrt{2\pi}\sigma_{1|2}} \exp[-\frac{1}{2}(\frac{Y_1 - \alpha_{1|2} - \beta_{12} Y_2}{\sigma_{1|2}})^2]
\end{aligned}
$$

Here

$$\alpha_{1|2} = \mu_1 - \mu_2 \rho_{12} \frac{\sigma_1}{\sigma_2}$$

$$\beta_{12} = \rho_{12} \frac{\sigma_1}{\sigma_2}$$

$$\sigma^2_{1|2} = \sigma^2_1 (1 - \rho^2_{12})$$

Inferences on Correlation Coefficients

Maximum Likelihood Estimation of ρ_{12} (Pearson Product-moment Correlation Coefficient):

$$r_{12} = \frac{\sum(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{[\sum(Y_{i1} - \bar{Y}_1)^2 \sum(Y_{i2} - \bar{Y}_2)^2]^{1/2}}$$

Usually biased, but bias is small when n is large.
Test whether $\rho_{12} = 0$

$$H_0 : \rho_{12} = 0$$

v.s.

$$H_1 : \rho_{12} \neq 0$$

Test Statistics:

$$t^* = \frac{r_{12}\sqrt{n-2}}{\sqrt{1 - r_{12}^2}}$$

If H_0 holds, t^* follows the $t(n-2)$ distribution.

Interval Estimation of ρ_{12}

Make the *Fisher z transformation*:

$$z' = \frac{1}{2}\log_e\left(\frac{1+r_{12}}{1-r_{12}}\right)$$

When n is large ($n > 25$), approximately normally distributed with

$$E\{z'\} = \zeta = \frac{1}{2}\log_e\left(\frac{1+\rho_{12}}{1-\rho_{12}}\right)$$

$$\sigma^2\{z'\} = \frac{1}{n-3}$$

Then we can make interval estimate

$$\frac{z' - \zeta}{\sigma\{z'\}}$$

is approximately standard normal. Then $1-\alpha$ confidence limits for ζ are

$$z' \pm z(1-\alpha/2)\sigma\{z'\}$$

Spearman Rank Correlation Coefficient

- Denote the rank of Y_{i1} by R_{i1} and the rank of Y_{i2} by R_{i2}. The ranks are from 1 to n.
- The Spearman Rank Correlation Coefficient r_S is then defined as

$$r_S = \frac{\sum(R_{i1} - \bar{R}_1)(R_{i2} - \bar{R}_2)}{[\sum(R_{i1} - \bar{R}_1)^2 \sum(R_{i2} - \bar{R}_2)^2]^{1/2}}$$

- as before $-1 \leq r_S \leq 1$.
- More robust compared with the Pearson correlation coefficient.