

# Inference in Normal Regression Model

Dr. Frank Wood

## Remember

- ▶ We know that the point estimator of  $b_1$  is

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

- ▶ Last class we derived the sampling distribution of  $b_1$ , it being  $N(\beta_1, \sigma^2\{b_1\})$  (when  $\sigma^2$  known) with

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

- ▶ And we suggested that an estimate of  $\sigma^2\{b_1\}$  could be arrived at by substituting the MSE for  $\sigma^2$  when  $\sigma^2$  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,  $(b_1 - \beta_1)/\sigma\{b_1\}$  is a standard normal variable  $N(0, 1)$
- ▶ We don't know  $\sigma^2\{b_1\}$  so it must be estimated from data. We have already denoted its estimate  $s^2\{b_1\}$
- ▶ Using this estimate we showed that

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

where

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

It is from this fact that our confidence intervals and tests will derive.

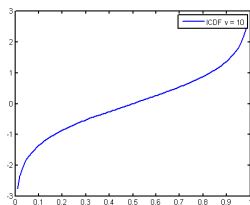
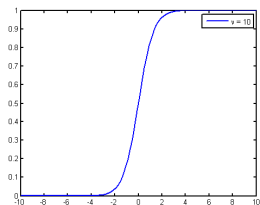
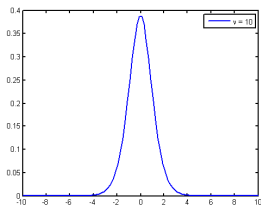
## Confidence Intervals and Hypothesis Tests

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

## Confidence Interval for $\beta_1$

Since the “studentized” statistic follows a t distribution we can make the following probability statement

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



matlab: tpdf, tcdf, tinu

## 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.  $\int_{-\infty}^y f(t)dt = p$

## Book tables and Matlab commands

Appendix B (or elsewhere in other books), a table of percentiles of the  $t$  distribution is given. In this table one number appears for each of a number of degrees of freedom  $\nu$  and a parameter, call it  $A$ .

Each entry is some value of  $t(A; \nu)$  where  $P\{t(\nu) \leq t(A; \nu)\} = A$

In words  $t(A; \nu)$  is the point on the horizontal axis of the Student- $t$  distribution where  $A$  percent of the mass under the curve is located to the left. This is precisely the quantity returned by  $\text{tinv}(A, \nu)$  in Matlab.

How can this be used to produce a confidence interval?

## Interval arriving from picking $\alpha$

- ▶ Note that by symmetry

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

- ▶ Remember

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

- ▶ Rearranging terms and using this symmetry we have

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

- ▶ And now we can use a table to look up and produce confidence intervals



## Using tables for Computing Intervals

- ▶ The tables in the book (table B.2 in the appendix) for  $t(1 - \alpha/2; \nu)$  where  $P\{t(\nu) \leq t(1 - \alpha/2; \nu)\} = A$
- ▶ Provides the inverse CDF of the t-distribution
- ▶ This can be arrived at computationally as well  
Matlab:  $\text{tinv}(1 - \alpha/2, \nu)$

## $1 - \alpha$ confidence limits for $\beta_1$

- ▶ The  $1 - \alpha$  confidence limits for  $\beta_1$  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  $\alpha$ .
  - ▶ Fixing  $\alpha$  can guide the choice of sample size if a particular confidence interval is desired
  - ▶ Give a sample size, vice versa.
- ▶ Also useful for hypothesis testing

# Tests Concerning $\beta_1$

- ▶ 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 $\beta_1$

- ▶ 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)

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

- ▶ Variability in  $b_1$  is assumed to arise from sampling noise.

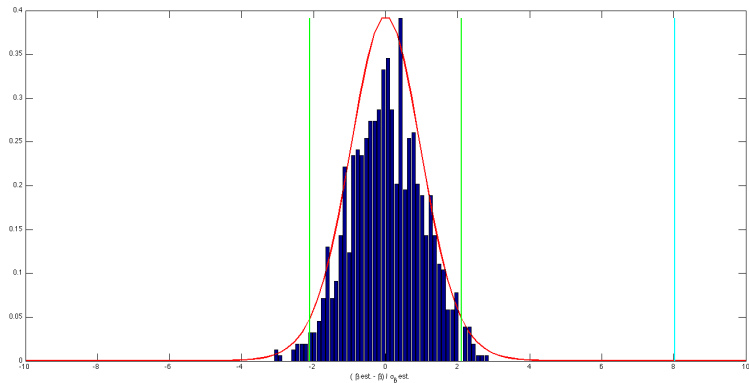
## Decision rules

if  $|t^*| \leq t(1 - \alpha/2; n - 2)$ , conclude  $H_0$

if  $|t^*| > t(1 - \alpha/2; n - 2)$ , conclude  $H_\alpha$

Absolute values make the test two-sided

# Intuition



p-value is value of  $\alpha$  that moves the green line to the blue line

## Calculating the p-value

- ▶ The p-value, or attained significance level, is the smallest level of significance  $\alpha$  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.
- ▶ In Matlab  
Two-sided p-value  
 $2 * (1 - tcdf(|t^*|, \nu))$

## Inferences Concerning $\beta_0$

- ▶ Largely, inference procedures regarding  $\beta_0$  can be performed in the same way as those for  $\beta_1$
- ▶ Remember the point estimator  $b_0$  for  $\beta_0$

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



## 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

$$E\{b_0\} = \beta_0$$

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

- ▶ When error variance is unknown

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

## Confidence interval for $\beta_0$

The  $1 - \alpha$  confidence limits for  $\beta_0$  are obtained in the same manner as those for  $\beta_1$

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

## Considerations on Inferences on $\beta_0$ and $\beta_1$

- ▶ Effects of departures from normality
  - ▶ The estimators of  $\beta_0$  and  $\beta_1$  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  $\sigma^2$ ) 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  $E\{Y_h\}$
- ▶ The point estimator of  $E\{Y_h\}$  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

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

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

## Sampling Distribution of $\hat{Y}_h$

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

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

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

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

when the  $Y_i$  are independent



## Sampling Distribution of $\hat{Y}_h$

Using this fact we have

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

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 the mean of  $Y$  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 (from last lecture)

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

- ▶ 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  $\sigma^2$  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 two 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 $E\{Y_h\}$

- ▶ The  $1 - \alpha$  confidence intervals for  $E\{Y_h\}$  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  $\hat{Y}_h$  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  $\hat{Y}_h$  reduces to the variance of the estimator  $b_0$  for  $\beta_0$



## Prediction interval for *new* input $X_h$

- ▶ Roughly the same idea as for  $E\{Y_h\}$  where  $X_h$  is a known input point included in the estimation of  $b_1, b_0$ , and  $s^2$
- ▶ If all regression parameters are known then the  $1 - \alpha$  prediction interval for a new observation  $Y_h$  is

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

## Prediction interval for *new* input $X_h$

- ▶ If the regression parameters are unknown the  $1 - \alpha$  prediction interval for a new observation  $Y_{h(new)}$  is given by the following theorem

$$\frac{Y_{h(new)} - \hat{Y}_h}{s\{\text{pred}\}} \sim t(n - 2)$$

for the normal error regression model.  $s\{\text{pred}\}$  to be defined shortly.

It follows directly that the  $1 - \alpha$  prediction limits for  $Y_{h(new)}$  are

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

- ▶ 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 *new* input $X_h$

Because  $Y_{h(\text{new})}$  is independent of  $\hat{Y}_h$  we can directly write

$$\sigma^2\{\text{pred}\} = \sigma^2\{Y_{h(\text{new})} - \hat{Y}_h\} = \sigma^2\{Y_{h(\text{new})}\} + \sigma^2\{\hat{Y}_h\} = \sigma^2 + \sigma^2\{\hat{Y}_h\}$$

where from before we have that

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

so

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

but as before we don't know  $\sigma^2$  so we will replace it...

## Prediction interval for *new* input $X_h$

The value of  $s^2\{pred\}$  is given by

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

Note that this quantity is *slightly* larger than  $s^2\{\hat{Y}_h\}$ .

It has two components

- ▶ The variance of the distribution of  $y$  at  $X = X_h$ , namely  $\sigma^2$
- ▶ The variance of the sampling distribution of  $\hat{Y}_h$ , namely  $s^2\{\hat{Y}_h\}$ .

## Summary

After this lecture you should be able to confidently do estimation, prediction, and hypothesis testing about the slope, intercept, and predicted values at any input point, old or new in the normal error linear regression setting.