# 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 $N(\beta_1, Var(b_1))$(when $\sigma^2$ 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 $\sigma^2$ 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 distribute, $(b_1 - \beta_1)/\sigma\{b_1\}$ is a standard normal variable $N(0,1)$
- We don't know $\text{Var}(b_1)$ so it must be estimated from data. We have already denoted it's estimate $s^2\{b_1\}$
- Using this estimate we it can be shown 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.

# Where does this come from?

- We need to rely upon (but will not derive) 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 $b_1$.

- Here there are two linear constraints

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

  imposed by the regression parameter estimation that each reduce the number of degrees of freedom by one (total two).

# Reminder: normal (non-regression) estimation

- Intuitively the regression result from the previous slide follows the standard result for the sum of squared standard normal random variables. First, with $\sigma$ and $\mu$ known

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left( \frac{Y_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

and then with $\mu$ unknown

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \\ \frac{(n-1)S^2}{\sigma^2} &= \sum_{i=1}^{n} \left( \frac{Y_i - \bar{Y}}{\sigma} \right)^2 \sim \chi^2(n-1) \end{aligned}$$

and $\bar{Y}$ and $S^2$ are independent. [1]

# Reminder: normal (non-regression) estimation cont.

- With both $\mu$ and $\sigma$ unknown then

$$\sqrt{n}\left(\frac{\bar{Y}-\mu}{S}\right) \sim t(n-1)$$

because

$$T = \frac{Z}{\sqrt{W/\nu}} = \frac{\sqrt{n}(\bar{Y}-\mu)/\sigma}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}} = \sqrt{n}\left(\frac{\bar{Y}-\mu}{S}\right)$$

Here the numerator follows from

$$Z = \frac{\bar{Y}-\mu_{\bar{Y}}}{\sigma_{\bar{Y}}} = \frac{\bar{Y}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$$

# Another useful fact : Student-t distribution

Let $Z$ and $\chi^2(\nu)$ be independent random variables (standard normal and $\chi^2$ respectively). We then define a t random variable as follows:

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

This version of the t distribution has one parameter, the degrees of freedom $\nu$

# Distribution of the studentized statistic

To derive the distribution of the statistic $\frac{b_1-\beta_1}{s\{b_1\}}$ first we do the following rewrite

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

where

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

# Studentized statistic cont.

And note the following

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

where we know (by the given theorem) the distribution of the last term is $\chi^2$ 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

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

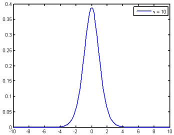
# Confidence Intervals and Hypothesis Tests

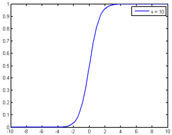
Now that we know the sampling distribution of $b_1$ (t with n-2 degrees of freedom) we can construct confidence intervals and hypothesis tests easily

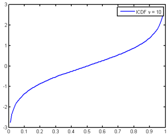
# Confidence Interval for $\beta_1$

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

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







# Remember

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

# Interval arriving from picking $\alpha$

- Note that by symmetry

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

- Rearranging terms and using this fact 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: $tinv(1-\alpha/2,\nu)$

# $1-\alpha$ 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)

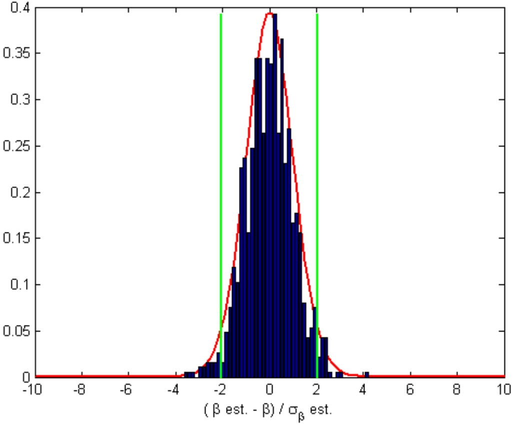
$$t^* = \frac{b_1 - 0}{s\{b_1\}}$$

# Decision rules

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

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

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

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

- 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 $\beta_0$

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 $\mathbb{E}(Y_h)$
- 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 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 $\mathbb{E}(Y_h)$

- 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 $\beta_0$

# 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

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

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