

# STA302/1001 - Methods of Data Analysis I

(Week 03 lecture notes)

Wei (Becky) Lin

Sept. 26, 2016



## Last Week

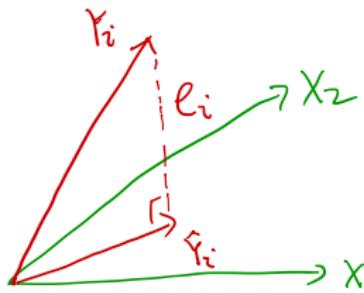
- 7 Properties of Least Squares Fitted Line.
- Why squared residuals instead of absolute residuals in OLS estimation?
- Introduction to other types of regressions
- Normal regression model.
- Maximum Likelihood Estimation (MLE) of  $\beta_0, \beta_1$ .
- Review of statistical inference.

## Week 03- Learning objectives & Outcomes

- Review of distribution theory.
- Inference for SLR.
- Interval estimation of mean response.
- Prediction interval.
- Difference between prediction interval and confidence interval.
- ANOVA approach

## Properties of LS fitted line (Summary)

0.  $E(\hat{Y}_i) = E(Y_i)$
  1.  $\sum_1^n e_i = 0$ .
  2.  $\sum_1^n e_i^2$  is a minimum.
  3.  $\sum_1^n Y_i = \sum_1^n \hat{Y}_i$
  4.  $\sum_1^n X_i e_i = 0$
  5.  $\sum_1^n \hat{Y}_i e_i = 0$
  6. The fitted regression line always goes through the point  $(\bar{X}, \bar{Y})$ .
- $\hat{Y}_i \perp\!\!\!\perp e_i$      $\perp\!\!\!\perp = \text{indep.}$



## Review of Distribution Theory

# Review of Distribution Theory

1. If  $X \sim N(\mu, \sigma^2)$ , then  $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$ .
2. If  $X_i \sim N(\mu_i, \sigma_i^2)$  independently, then

$$W = a + \sum_1^n a_i X_i \sim N(a + \sum_1^n a_i \mu_i, \sum_1^n a_i^2 \sigma_i^2)$$

3. If  $Z_i \sim_{iid} N(0, 1)$ , then  $\sum_1^n Z_i^2 \sim \chi_n^2$

4. If  $Z \sim N(0, 1) \perp X \sim \chi_n^2$ , then  $T = \frac{Z}{\sqrt{X/n}} \sim t_n$ .

5. If  $X_1 \sim \chi_n^2 \perp X_2 \sim \chi_m^2$ , then  $\frac{X_1/n}{X_2/m} \sim F_{n,m}$

6. If  $X_1, \dots, X_n \sim_{iid} N(\mu, \sigma^2)$  then

- $\bar{X} \sim N(\mu, \sigma^2/n)$

- $s^2 = \sum(X_i - \bar{X})^2/(n-1)$ , and  $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$

sample  
variance,

$$\mathbb{E}(s^2) = \sigma^2$$

$Z$  is the standardized  $X$

$$T^* = \frac{N(0,1)}{\sqrt{\chi_n^2/n}} \sim t_n$$

$$F^* = \frac{\chi_n^2/n}{\chi_m^2/m} \sim F_{n,m}$$

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

## Chapter 2

# Inference in Regression and Correlation Analysis

## Assume Normal Error model

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

where

- $\beta_0, \beta_1$  are regression coefficients or model parameters.
- $X_i$  are known constants.
- $\epsilon_i \sim_{iid} N(0, \sigma^2)$ .

Recall • Gauss-Markov Conditions on  $\epsilon_i$

- {  
• For RUs  $X, Y$

$$\text{cov}(X, Y) = 0$$

$$\overbrace{\qquad\qquad\qquad}^{X \perp\!\!\!\perp Y}$$

$$\left\{ \begin{array}{l} \text{① } E(\epsilon_i) = 0 \\ \text{② } V(\epsilon_i) = \sigma^2 \\ \text{③ } \text{cov}(\epsilon_i, \epsilon_j) = 0 \end{array} \right.$$

# Sampling distribution of $b_1$

$$\varepsilon_i \sim \text{iid } N(0, \sigma^2)$$

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

$$\Rightarrow \text{Var}(Y_i) = \text{Var}(\varepsilon_i) = \sigma^2$$

$$\Rightarrow \text{Var}(b_1) = \text{Var}(\sum_i^n k_i Y_i)$$

- From week 1 lecture, we know

- $b_1 = \frac{s_{XY}}{S_{XX}} = \sum_1^n k_i Y_i$ , so the  $b_1$  is normally distributed.

- $b_1$  is BLUE.  $E(b_1) = \beta_1, \sigma^2(b_1) = V(b_1) = \frac{\sigma^2}{S_{XX}}$

- $\hat{\sigma}^2 = \text{MSE} = \text{SSE}/(n-2)$ , so  $s^2(b_1) = \text{MSE}/S_{XX}, \sqrt{s^2(b_1)} = s(b_1)$

- Sampling distribution of  $b_1$  Normal error model assumption

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

$V(b_1) = \sigma^2 = \frac{\sigma^2}{S_{XX}}$

$\hat{V}(b_1) = S^2(b_1) = \frac{\text{MSE}}{S_{XX}}$

- $\frac{b_1 - \beta_1}{s(b_1)} / \sqrt{\frac{s^2(b_1)}{\sigma^2(b_1)}}$

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

- $\frac{s^2(b_1)}{V(b_1)} = \frac{\text{MSE}}{S_{XX}} / \frac{\sigma^2}{S_{XX}} = \left( \frac{\text{MSE}}{\sigma^2} \right) = \frac{\chi^2_{n-2}}{n-2}$

- $\frac{(n-2)\text{MSE}}{\sigma^2} \sim \chi^2_{n-2}$

$$\frac{b_1 - \beta_1}{s(b_1)} = \frac{b_1 - \beta_1}{\sigma(b_1)} \div \sqrt{\frac{s^2(b_1)}{\sigma^2(b_1)}}$$

$$= N(0, 1) / \sqrt{\frac{\chi^2_{n-2}}{n-2}}$$

$$\sim t_{n-2}$$

## Inferences concerning $\beta_1$

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

$$b_1 \pm t_{1-\alpha/2, n-2} s(b_1) = b_1 \pm t_{1-\alpha/2, n-2} \sqrt{\frac{MSE}{S_{xx}}}$$

- Testing concerning  $\beta_1$

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

Note that when  $\beta_1 = 0$ , there is no linear association between  $Y$  and  $X$ .

- Test statistics

$$t^* = \frac{b_1 - 0}{s(b_1)} \mid_{H_0} \sim t_{n-2}$$

- If  $|t^*| \leq t_{1-\alpha/2, n-2}$ , fail to reject  $H_0$ .
- If  $|t^*| > t_{1-\alpha/2, n-2}$ , reject  $H_0$ .
- Or find relevant p-value, and we reject  $H_0$  if p-value is less than  $\alpha$ .
- If we change  $H_a : \beta > 0$  then we reject  $H_0$  if  $t^* > t_{1-\alpha, n-2}$  otherwise we fail to reject null hypothesis.

# Sampling distribution of $b_0$

$$V(b_0) = \sigma^2(b_0) = V(\sum_i^n w_i Y_i)$$

$$= \sigma^2 \underbrace{\sum_i^n w_i^2}_{\bar{x}^2}$$

$$= \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

- From week 1 lecture, we know

- $b_0 = \bar{Y} - b_1 \bar{X} = \sum_1^n w_i Y_i$ , so the  $b_0$  is normally distributed.
- $b_0$  is BLUE.  $E(b_0) = \beta_0$ ,  $\sigma^2(b_0) = V(b_0) = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right\}$
- $s^2(b_0) = MSE \left\{ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right\}$

- Sampling distribution of  $b_0$  under Normal error model assumption

$$\frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2}$$

$$\begin{cases} V(b_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \\ S^2(b_0) = MSE \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \end{cases}$$

- $\frac{b_0 - \beta_0}{\sigma(b_0)} / \sqrt{\frac{s^2(b_0)}{\sigma^2(b_0)}}$
- $\frac{b_0 - \beta_0}{\sigma(b_0)} \sim N(0, 1)$
- $\frac{s^2(b_0)}{V(b_0)} = \frac{MSE}{S_{xx}} / \frac{\sigma^2}{S_{xx}} = \frac{MSE}{\sigma^2} \sim \frac{\chi^2_{n-2}}{n-2}$
- $\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$

$$\begin{aligned} \frac{b_0 - \beta_0}{s(b_0)} &= \frac{b_0 - \beta_0}{\sigma(b_0)} \div \sqrt{\frac{s^2(b_0)}{\sigma^2(b_0)}} \\ &= N(0, 1) / \sqrt{\frac{\chi^2_{n-2}}{(n-2)}} \\ &\sim t_{n-2} \end{aligned}$$

## Inferences concerning $\beta_0$

There are only infrequent occasions when we wish to make inferences concerning  $\beta_0$ . We do inferences concerning  $\beta_0$  only when the scope of the model includes  $X = 0$ .

- $1 - \alpha$  confidence limits for  $\beta_0$  are

$$b_0 \pm t_{1-\alpha/2, n-2} s(b_0) = b_0 \pm t_{1-\alpha/2, n-2} \sqrt{MSE \left\{ \frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} \right\}}$$

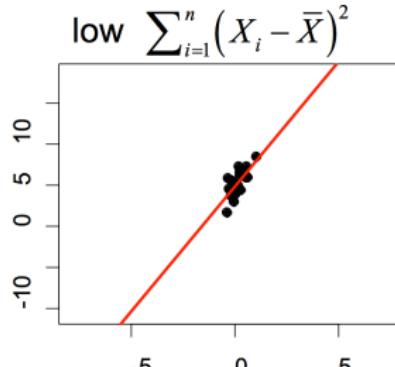
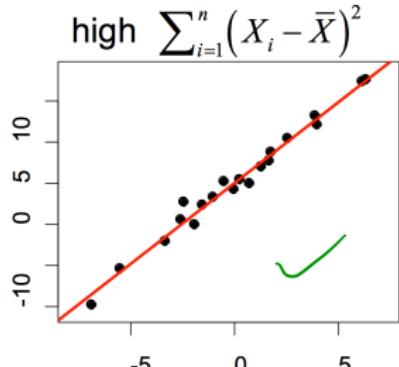
- Testing concerning  $\beta_0$  is less interesting.

## Dicussion: sampling distribution of $b_0, b_1$

$$V(b_1) = \frac{\sigma^2}{S_{XX}} = \frac{\sigma^2}{\sum_i(X_i - \bar{X})^2}$$

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

- $V(b_1), V(b_0)$  both are inversely related to  $S_{XX} = \sum_i(X_i - \bar{X})^2$ .
- Higher  $S_{XX}$ , lower variance.  $\Rightarrow$  narrow confidence interval for  $\beta_0, \beta_1$
- Which estimate do you trust more?



## Example 2 (continued from last week)

### Annual Salary (Y) vs years of service (X)

```
X=c(3,6,10,8,13)      # assign predictor observations to object X
Y=c(34,34,38,37,47)  # assign response observations to object Y
lmfit = lm(Y~X)       # fitting data with a simple linear regression
# summary(lmfit)      # summary of the fitted model
```

```
##  
## Call:  
## lm(formula = Y ~ X)  
##  
## Residuals:  
##      1      2      3      4      5  
##  2.293 -1.483 -2.517 -1.000  2.707  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 27.9310 =b0 3.1002 =s(b0) 9.01 0.00289 **  
## X          1.2586 =b1  0.3566 =s(b1) 3.53 0.03864 *  
## ---  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 2.715 on 3 degrees of freedom  
## Multiple R-squared: 0.806, Adjusted R-squared: 0.7413  
## F-statistic: 12.46 on 1 and 3 DF, p-value: 0.03864  
= 3.53^2
```

H0:  $\beta_1 = 0$   
Ha:  $\beta_1 \neq 0$   
 $\frac{1.2586}{0.3566} = 3.53$   
 $t = \frac{b_1}{s(b_1)} = 3.53$   
 $P(t > 3.53) = 2 * P(t_3 > 3.53)$

# Interval Estimation of $E(Y_h)$

- Let  $X_h$  denote the level of X for which we wish to estimate the mean response.
- The mean response when  $X = X_h$  is denoted by  $E(Y_h) = \beta_0 + \beta_1 X_h$
- Point estimate of  $E(Y_h)$  is  $\hat{Y}_h = b_0 + b_1 X_h$ .
- $E(\hat{Y}_h) = E(b_0 + b_1 X_h) = \beta_0 + \beta_1 X_h = E(Y_h)$
- Show  $V(\hat{Y}_h) = \sigma^2 \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}} \right\}$
- Proof:
  - $Cov(\bar{Y}, b_1) = 0$  (Practise problem last week)
  -

$$\begin{aligned} V(\hat{Y}) &= V\left(\frac{1}{n} \sum_i^n Y_i\right) \quad V(\hat{Y}_h) = V(b_0 + b_1 X_h) = V\{\bar{Y} + b_1(X_h - \bar{X})\} \\ &= \left(\frac{1}{n}\right)^2 \sum_i^n V(Y_i) \quad = V(\bar{Y}) + (X_h - \bar{X})^2 V(b_1) + 2(X_h - \bar{X}) \text{Cov}(\bar{Y}, b_1) \\ &= \frac{1}{n^2} \cdot n \sigma^2 \quad = \sigma^2 \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}} \right\} \quad \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} = 0 \\ &= \sigma^2/n \end{aligned}$$

- When MSE is substituted, the estimated variance of  $\hat{Y}_h$   
 $s^2(\hat{Y}_h) = \text{MSE} \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}} \right\}$   
 $s^2(\hat{Y}_h)$  is an unbiased estimator of  $V(\hat{Y}_h)$
- When  $X_h = 0$ ,  $\hat{Y}_h$  reduces to  $b_0$  and  $s^2(\hat{Y}_h)$  is  $s^2(b_0)$ .

# Interval Estimation of $E(Y_h)$

- Sampling distribution of  $\hat{Y}_h$

$$\left| \frac{\hat{Y}_h - E(\hat{Y}_h)}{s(\hat{Y}_h)} \right| \sim t_{n-2}$$

Proof:

$$\left| \frac{\hat{Y}_h - E(\hat{Y}_h)}{\sigma(\hat{Y}_h)} \div \sqrt{\frac{s^2(\hat{Y}_h)}{\sigma^2(\hat{Y}_h)}} \right| = N(0,1) \div \sqrt{\frac{\chi^2_{n-2}}{n-2}} \sim t_{n-2}$$

where

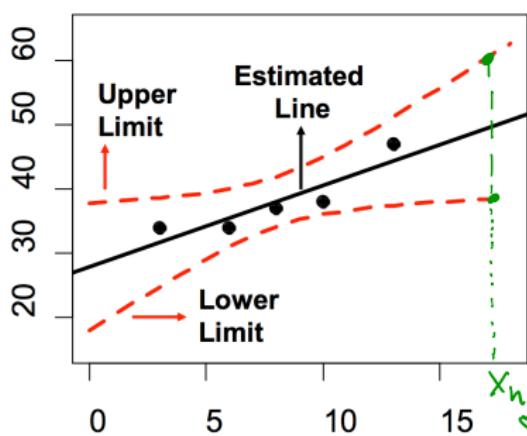
$$\frac{\hat{Y}_h - E(\hat{Y}_h)}{\sigma(\hat{Y}_h)} \sim N(0,1), \quad \frac{s^2(\hat{Y}_h)}{\sigma^2(\hat{Y}_h)} \sim \frac{\chi^2_{n-2}}{n-2}$$

- $1 - \alpha$  confidence interval for  $E(Y_h)$

$$\hat{Y}_h \pm t_{1-\alpha/2, n-2} s(\hat{Y}_h) = \hat{Y}_h \pm t_{1-\alpha/2, n-2} \sqrt{MSE\left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}} \right\}}$$

# Confidence Interval for $E(Y_h)$

- Plot of Confidence limits for  $E(Y_h)$  at different level of  $X_h$ .



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

The further  $X_h$  is from  $\bar{X}$ ,  
the higher the uncertainty  
and, thus, the wider the CI

This level of  $X_h$  is far from  $\bar{X}$

$\Rightarrow s^2(\hat{Y}_h)$  is larger

$\Rightarrow$  CI for  $E(Y_h)$  at  $X=X_h$   
is wider.

## Prediction Intervals for $Y_{h(\text{new})}$

↳ a random variable

Assume normal error regression model

Estimate it by  $E(Y_{h(\text{new})})$

Use  $\hat{Y}_h$

- $Y_{h(\text{new})} = \beta_0 + \beta_1 X_h + \epsilon_h$
- Assume  $\beta_0, \beta_1, \sigma^2$  are known. Then the  $1-\alpha$  prediction limits for  $Y_{h(\text{new})}$  is  
=) uncertainty comes only from  $\epsilon_h; \sigma^2$

$$E(Y_{h(\text{new})}) \pm Z_{1-\alpha/2} \sigma, \quad E(Y_{h(\text{new})}) = \beta_0 + \beta_1 X_h$$

- If model parameters are unknown:

- $\hat{Y}_h = b_0 + b_1 X_h$  is the point estimate of  $E(Y_{h(\text{new})}) = \beta_0 + \beta_1 X_h$
- $Y_{h(\text{new})} \perp \hat{Y}_h$  since a new observation and the original  $n$  cases on which  $\hat{Y}_h$  is based are independent.
- $\sigma_{\text{pred}}^2 = V(Y_{h(\text{new})} - \hat{Y}_h) = V(Y) + V(\hat{Y}_h) = \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}} \right\}$  which has two components \*
- The variance of the distribution of  $Y$  at  $X = X_h$ , namely  $\sigma^2$
- The variance of the possible location of the distribution  $Y$ , i.e., the variance of the sampling distribution of  $\hat{Y}_h$ , namely,  $V(\hat{Y}_h)$ .

L or II:  
indep.

## Prediction Intervals for $Y_{h(\text{new})}$

estimate

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

- Replace the  $\sigma^2$  with MSE, an unbiased estimator of  $\sigma_{\text{pred}}^2$  is

$$s_{\text{pred}}^2 = \text{MSE} \left\{ 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right\}$$

- The  $1-\alpha$  prediction limits for a single new observation  $Y_{h(\text{new})}$  are

$$\hat{Y}_h \pm t_{1-\alpha/2, n-2} s_{\text{pred}}$$

## Example: Confidence intervals vs Prediction intervals



## Confidence interval vs Prediction intervals

$\sigma_{\text{pred}}$

$\sigma(\hat{Y}_h)$

- The difference between a prediction interval and a confidence interval is the **standard error**.
- The standard error for a **confidence interval** on the mean **takes into account the uncertainty due to sampling**. The line you computed from your sample will be different from the line that would have been computed if you had the entire population, the standard error takes this uncertainty into account.
- The standard error for a **prediction interval** on an individual observation takes into account the uncertainty due to **sampling** like above, but also takes into account **the variability of the individuals around the predicted mean**. The se for the prediction interval will be wider than for the confidence interval and hence the prediction interval will be wider than the confidence interval
- A **prediction interval** is **wider** than a **confidence interval** for the same  $\alpha$  level.

## Confidence interval vs Prediction intervals

- Prediction interval is wider the further  $X_h$  is from  $\bar{X}$ . This is due to the fact that  $V(\hat{Y}_h)$  goes up as  $X_h - \bar{X}$  increases.
- The prediction limits, unlike the confidence limits for  $E(Y_h)$ , are sensitive to departures from normality of the error term distribution. (coming later...)
- Prediction intervals resemble confidence intervals. However, they differ conceptually,
  - A CI represents an inference on a parameter and is an interval that is intended to cover the value of the parameter.
  - A PI, on the other hand, is a statement about the value to be taken by a random variable, the new observation  $Y_{h(new)}$ .

## Confidence Band for Regression line

- To obtain a confidence band for the entire regression line  $E(Y) = \beta_0 + \beta_1 X$ . This band enable us to see the region in which the entire regressin line lies.
- The Working-Hotelling  $1 - \alpha$  confidence band for the regression line

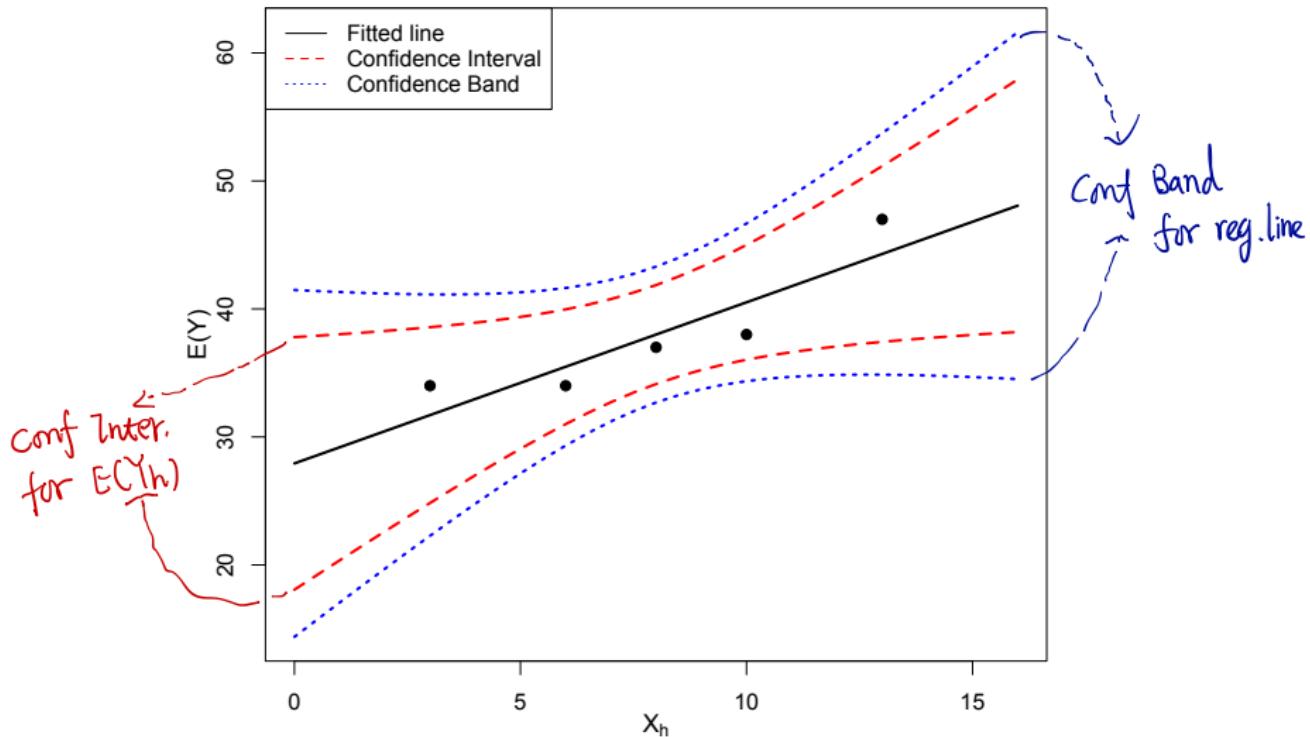
$$\hat{Y}_h \pm W s(\hat{Y}_h) = (b_0 + b_1 X_h) \pm \sqrt{MSE \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}} \right\}}$$

where  $W^2 = 2F_{1-\alpha;2,n-2}$

- Since we are doing all values of  $X_h$  at once, it will be wider at each  $X_h$  than CIs for individual  $E(Y_h)$ .
- The boundary values of the confidence band for the regression line at any value  $X_h$  often are not substantially wider than the confidence limits for the mean response  $E(Y_h)$  at  $X_h$ .

# Confidence Band vs Confidence Limits

Annual salary vs Years of service



# Confidence Band vs Confidence Limits vs Prediction Limits

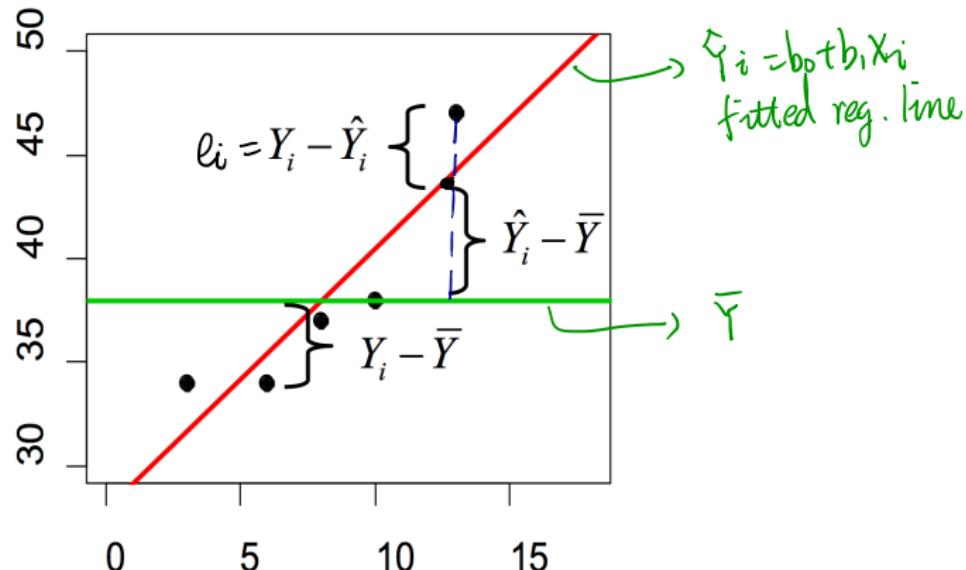


Analysis of Variance (ANOVA) approach  
to Regression Analysis

## Main idea of ANOVA

- ANOVA stands for Analysis of Variance.
- The approach analyzes variation in the data
  - Partitioning of Sum of Squares (SS) into various sources.
  - Compare SS's and make inferences for model and for various parameters.
- This approach is quite useful in multiple linear regression (more than one predictor)

## ANOVA: SS breakdown



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

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

$\underbrace{\quad}_{SST}$        $\underbrace{\quad}_{SSE}$        $\underbrace{\quad}_{SSR}$

## ANOVA: SS breakdown

- Total sum of squares

$$SST = \sum_i (Y_i - \bar{Y})^2$$

- SSE: the error sum of squares

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

- SSR: the regression sum of Squares

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

## ANOVA: $SST = SSR + SSE$

$$\begin{aligned} SST &= \sum_{i=1}^n (Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^n \{(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})\}^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\ &= \underbrace{\sum_{i=1}^n e_i^2}_{SSE} + \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{SSR} + 2 \sum_{i=1}^n e_i (\hat{Y}_i - \bar{Y}) \\ &= SSE + SSR + 2 \underbrace{\sum_{i=1}^n e_i \hat{Y}_i}_{=0} - 2\bar{Y} \underbrace{\sum_{i=1}^n e_i}_{=0} \\ &= SSE + SSR \end{aligned}$$

Q.E.D.

ANOVA: show  $SSR = b_1^2 \sum_1^n (X_i - \bar{X})^2$

$$\begin{aligned} SSR &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 && \downarrow \hat{Y}_i = b_0 + b_1 X_i \\ &= \sum_{i=1}^n (\underline{b_0 + b_1 X_i} - \bar{Y})^2 && \curvearrowleft b_0 = \bar{Y} - b_1 \bar{X} \\ &= \sum_{i=1}^n (\underline{\bar{Y} - b_1 \bar{X}} + b_1 X_i - \bar{Y})^2 \\ &= \sum_{i=1}^n (b_1 (X_i - \bar{X}))^2 \\ &= b_1^2 \sum_1^n (X_i - \bar{X})^2 \end{aligned}$$

Q.E.D.

# ANOVA: Mean Squares

- Mean Squares are the SS's divided by their degrees of freedom. In SLR, we have:

Sum of Squares(SS)	Degree of Freedom(df)	Mean squares(MS)
SSR	1	$MSR = SSR/1$
SSE	$n-2$	$MSE = SSE/(n-2)$
SST	$n-1 = 1 + (n-2)$	-

- $df(SST) = df(SSR) + df(SSE)$
- MSR: Regression Mean Squares; MSE=Error Mean Squares.
- Properties of MS(\*)

- $E(MSE) = \sigma^2$
- $E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$
- $\frac{MSR}{\sigma^2} |_{\beta_1=0} \sim \chi_1^2$ , and  $\frac{(n-2)MSE}{\sigma^2} \sim \chi_{n-2}^2$

$$\text{ANOVA: } E(\text{MSR}) = \sigma^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\text{SSR} = \sum_i^n (\hat{Y}_i - \bar{Y})^2 = b_1^2 \sum_i^n (X_i - \bar{X})^2 = S_{xx} b_1^2$$

$$\text{MSR} = \text{SSR}/1$$

$$E(\text{MSR}) = E(\text{SSR}) = E(S_{xx} b_1^2)$$

$$= S_{xx} E(b_1^2)$$

$$= S_{xx} (V(b_1) + E(b_1)^2)$$

$$= S_{xx} \left( \frac{\sigma^2}{S_{xx}} + \beta_1^2 \right)$$

$$= \sigma^2 + \beta_1^2 S_{xx}$$

$$= \sigma^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$$

for a RV  $X$

$$\begin{aligned} \text{Var}(X) &= E(X^2) - (EX)^2 \\ \Rightarrow E(X^2) &= V(X) + (EX)^2 \end{aligned}$$

Q.E.D.

# ANOVA: F test of $\beta_1 = 0$ Versus $\beta_1 \neq 0$

$$H_0 : \beta_1 = 0, \quad \beta_0 \neq 0$$

$$\frac{MSR}{\sigma^2} \Big|_{H_0: \beta_1=0} \sim \chi^2_1, \quad E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$$
$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}, \quad E(MSE) = \sigma^2$$

$\left. \begin{array}{l} MSR > MSE : \beta_1 \neq 0 \\ MSR \approx MSE : \beta_1 = 0 \end{array} \right\}$

- $F^* = MSR/MSE$ , if  $MSR > MSE$ , large  $F^*$  support  $H_a$  and values of  $F^*$  near 1 support  $H_0$ . The appropriate test is an upper-tail one.
- Under  $H_0$ , we have

$$F^*|_{H_0} = \underbrace{\left\{ \frac{MSR}{\sigma^2} / 1 \right\}}_{\chi^2_1 / 1} \div \underbrace{\left\{ \frac{(n-2)MSE}{\sigma^2} / (n-2) \right\}}_{\chi^2_{n-2} / (n-2)} = \boxed{\frac{MSR}{MSE} \sim F_{1, n-2}}$$

- To test  $H_0$ 
  - Calculate  $F^*$
  - Reject  $H_0$  if  $F^* > F_{(1-\alpha; 1, n-2)}$

# ANOVA: equivalence of F test and t test

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

$$F^* = \frac{SSR \div 1}{SSE \div (n-2)} = \frac{b_1^2 \sum (X_i - \bar{X})^2}{MSE}$$

**Proof:**

- For testing:  $H_0: \beta_1 = 0$  vs  $H_a: \beta_1 \neq 0$

$$t^* = \frac{b_1}{S(b_1)} \Big|_{H_0} \sim t_{n-2}, \text{ where } S(b_1) = \sqrt{\frac{MSE}{S_{xx}}}$$

$$\begin{aligned} F^* &= \frac{b_1^2 \sum (X_i - \bar{X})^2}{MSE} = \frac{b_1^2}{MSE/S_{xx}} = \left( \frac{b_1}{\sqrt{MSE/S_{xx}}} \right)^2 \\ &= (t^*)^2 \sim [F_{1, n-2}] \end{aligned}$$

$$\therefore t^* = \frac{b_1 - 0}{\sigma(b_1)} / \sqrt{\frac{s^2(b_1)}{\sigma^2(b_1)}} = \frac{N(0, 1)}{\sqrt{\chi^2_{n-2}/(n-2)}} \Rightarrow (t^*)^2 = \frac{\chi^2_1/1}{\chi^2_{n-2}/(n-2)}$$

## ANOVA: F test and t test

- At any given  $\alpha$  level, we can use either the t test or the F test for  $\beta_1 = 0$  versus  $\beta_1 \neq 0$
- The t test, is more flexible since it can be used for one-sided alternatives involving  $\beta_1(\geq, \leq)0$  versus  $\beta(>, <)0$ , while F test cannot.

## Example 2: Annual Salary (Y) vs years of service (X)

```
X=c(3,6,10,8,13)      # assign predictor observations to object X  
Y=c(34,34,38,37,47) # assign response observations to object Y  
lmfit = lm(Y~X)       # fitting data with a simple linear regression  
# summary(lmfit)      # summary of the fitted model
```

```
##  
## Call:  
## lm(formula = Y ~ X)  
##  
## Residuals:  
##      1      2      3      4      5  
##  2.293 -1.483 -2.517 -1.000  2.707  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  27.9310 =b0  3.1002=s(b0) 9.01   0.00289 **  
## X           1.2586 =b1  0.3566=s(b1) 3.53    0.03864 *  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 2.715 on 3 degrees of freedom  
## Multiple R-squared:  0.806, Adjusted R-squared:  0.7413  
## F-statistic: 12.46 on 1 and 3 DF, p-value: 0.03864
```

b1/s(b1)

H0: beta1=0  
Ha: beta1 <> 0  
12\*P(t\_3>3.53)

=MSE^{1/2}

$$= 3.53^2, F^* = (t^*)^2$$

## Example 2: Annual Salary (Y) vs years of service (X)

```
X=c(3,6,10,8,13)      # assign predictor observations to object X  
Y=c(34,34,38,37,47)  # assign response observations to object Y  
lmfit = lm(Y~X)       # fitting data with a simple linear regression  
anova(lmfit)          # anova approach
```

```
## Analysis of Variance Table  
##  
## Response: Y  
##           Df Sum Sq Mean Sq F value Pr(>F)  
## X          1 91.879 91.879 12.461 0.03864 *## Residuals  3 22.121  7.374  
## ---  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$MSR = 91.879 / 1$$

$$F^* = \frac{MSR}{MSE} = \frac{91.879}{7.374}$$

SSR

$$MSE = 22.121 / 3$$

SSE

$$3.53^2$$

$$\text{## [1] } 12.4609$$

$$\text{double check } (t^*)^2 = F^*$$

## Practice problems and upcoming topics

- Practice problems after today's lecture: Chapter 2: 2.1, 2.2, 2.3, 2.5, 2.6 (skip part e), 2.8 (skip part b), 2.10, 2.12, 2.14, 2.15, 2.18
- Upcoming topics
  - correlation coefficient
  - sample correlation coefficient
  - Inference for Correlation Coefficient
  - Coefficient of determination
  - Interpretation of  $R^2$
  - Normal correlation model
- Reading for next week's lecture: CH2.8, 2.9, 2.11