STA302/1001 - Methods of Data Analysis I

(Week 02 - Lectue A)

Wei (Becky) Lin

May 22-26, 2017

Last Week

Remind you: A1 is due this Thursday, May 25, 11pm.
Late submission penalty: 10% perday

- 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 β_0, β_1.
- Review of statistical inference.

Week 02-Lecture A: Learning objectives & Outcomes

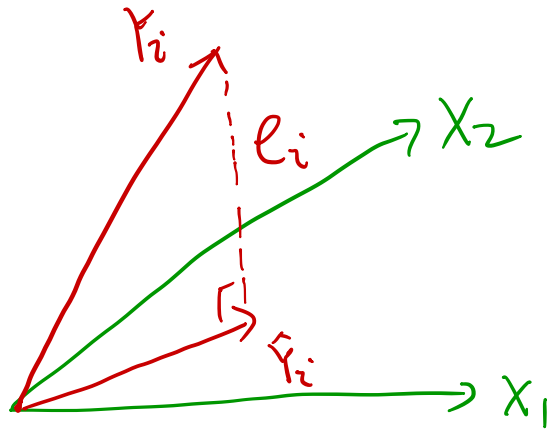
- Review of distribution theory.
- Inference for SLR.
- Interval estimation of mean response.
- Prediction inteval.
- Difference between prediction interval and confidence interval.
- ANOVA approach

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. $\sum_1^n \hat{Y}_i e_i = 0$

$\longrightarrow \hat{Y}_i \perp e_i$

$\perp$ = indep.



6. The fitted regression line always goes through the point $(\bar{X}, \bar{Y})$.

Review of Distribution Theory

Review of Distribution Theory

1. If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. Z is the standardized X
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 $X_1, \ldots X_n \sim_{iid} N(\mu, \sigma^2)$ then
   - $\bar{X} \sim N(\mu, \sigma^2/n)$
   - $s^2 = \sum(X_i - \bar{X})^2/(n-1)$, and $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$ *
   - $\bar{X} \perp s^2$, and

sample variance, $E(s^2) = \sigma^2$

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

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

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

Chapter 2
Inference in Regression and Correlation Analysis

Assume Normal Error model

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

where

- β_0, β_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 ε_i
① $E(\varepsilon_i) = 0$
② $V(\varepsilon_i) = \sigma^2$
③ $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$

• For RVs X, Y
$\mathrm{cov}(X, Y) = 0 \not\Rightarrow X \perp\!\!\!\perp Y$
$X \perp\!\!\!\perp Y \Rightarrow \mathrm{cov}(X, Y) = 0$

Sampling distribution of b_1

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

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

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

$\Rightarrow Var(b_1) = Var(\sum_1^n k_i Y_i)$

$= \sigma^2 \sum_1^n k_i^2 = \frac{\sigma^2}{S_{XX}}$

- 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. $E(b_1) = \beta_1, \sigma^2(b_1) = V(b_1) = \frac{\sigma^2}{S_{XX}}$
  - $\hat{\sigma}^2$=MSE=SSE/(n-2), so $s^2(b_1) = MSE/S_{XX}$, $\sqrt{s^2(b_1)} = s(b_1)$
- Sampling distribution of b_1 Normal error model assumption

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

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

- $\frac{b_1-\beta_1}{\sigma(b_1)} / \sqrt{\frac{s^2(b_1)}{\sigma^2(b_1)}}$
- $\frac{b_1-\beta_1}{\sigma(b_1)} \sim N(0,1)$
- $\frac{s^2(b_1)}{V(b_1)} = \frac{MSE}{S_{XX}} / \frac{\sigma^2}{S_{XX}} = \frac{MSE}{\sigma^2} = \sim \frac{\chi^2_{n-2}}{n-2}$
- $\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$

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

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

$\sim t_{n-2}$

Inferences concerning β_1

- $1-\alpha$ confidence limits for β_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 β_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

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

  - 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 α.
  - 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

$V(b_0) = \sigma^2(b_0) = V(\sum_1^n w_i Y_i)$

$= \sigma^2 \sum_1^n w_i^2$

$= \sigma^2 (\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}})$

- 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. $E(b_0) = \beta_0, \sigma^2(b_0) = V(b_0) = \sigma^2\{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\}$
  - $s^2(b_0) = MSE\{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\}$
- Sampling distribution of b_0 under Normal error model assumption

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

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

$s^2(b_0) = MSE(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}})$

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

$\frac{b_0 - \beta_0}{s(b_0)} = \frac{b_0 - \beta_0}{\sigma(b_0)} \div \sqrt{\frac{s^2(b_0)}{\sigma^2(b_0)}}$

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

$\sim t_{n-2}$

Inferences concerning β_0

There are only infrequent occasions when we wish to make inferences concerning β_0. We do inferences cencerning β_0 only when the scope of the model includes $X = 0$.

- $1 - \alpha$ confidence limits for β_0 are

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

- Testing concerning β_0 is less interest.

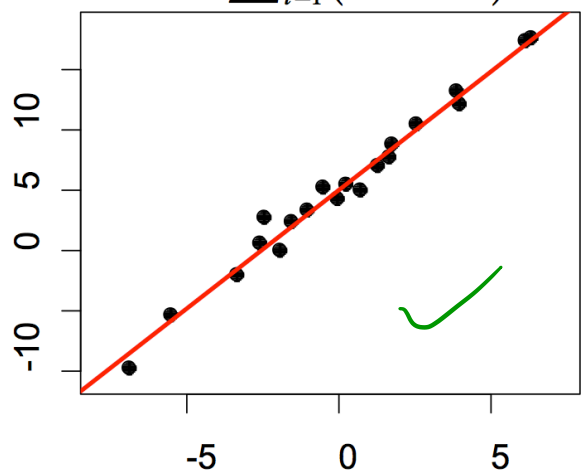
Dicussion: sampling distribution of b_0, b_1

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

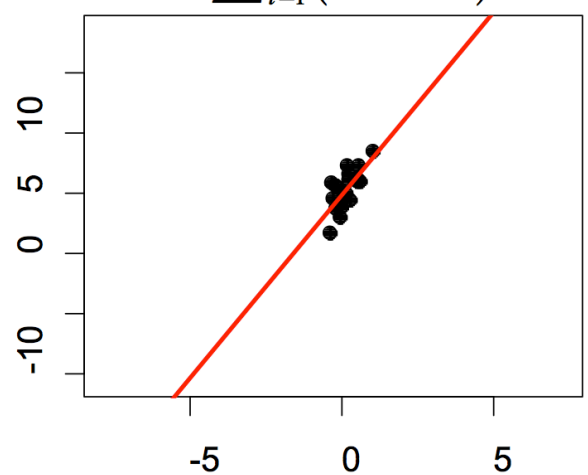
$$V(b_0) = \sigma^2\{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\} = \sigma^2\{\frac{1}{n} + \frac{\bar{X}^2}{\sum_i (X_i - \bar{X})^2}\}$$

- $V(b_1), V(b_0)$ both are inversely related to $S_{XX} = \sum_i (X_i - \bar{X})^2$.
- Higher S_{XX}, lower variance. ⇒ narrow confidence Interal. for β_0, β_1
- Which estimate do you trust more?

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



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



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
```

1.2586/0.3566

11

b1/s(b1)

H0: beta1=0

Ha: beta <>0

12*P(t_3>3.53)

=MSE^{1/2}

=3.53^2

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 $E(Y_h)$
- 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 $V(\hat{Y}_h) = \sigma^2\{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}\}$
- Proof:
  - $Cov(\bar{Y}, b_1) = 0$ (Practise problem last week)
  -

$$V(\hat{Y}_h) = V(b_0 + b_1 X_h) = V\{\bar{Y} + b_1(X_h - \bar{X})\}$$
$$= V(\bar{Y}) + (X_h - \bar{X})^2 V(b_1) + 2(X_h - \bar{X}) Cov(\bar{Y}, b_1)$$
$$= \sigma^2\{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}\}$$

- When MSE is substituted, the estimated variance of $\hat{Y}_h$

$$s^2(\hat{Y}_h) = MSE\{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}\}$$

- 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 $\hat{Y}_h$

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

Proof:

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

where

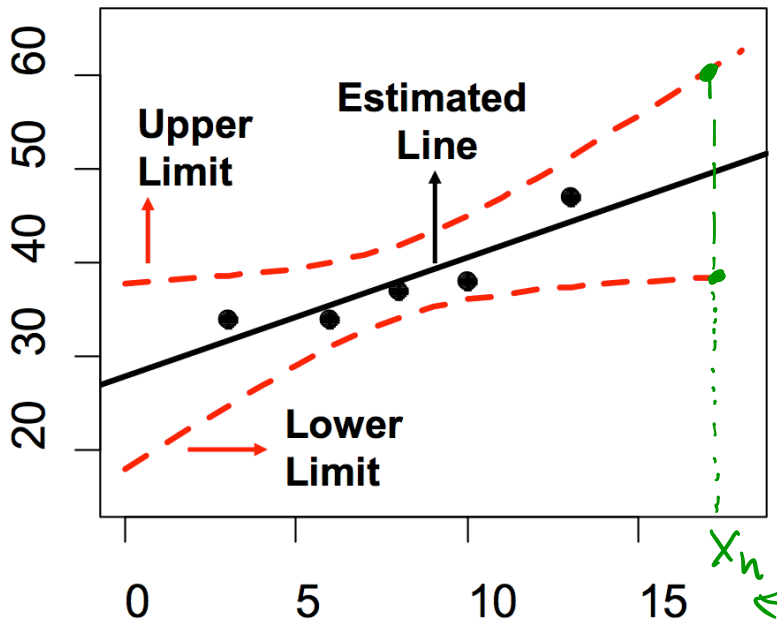
$$\frac{\hat{Y}_h - E(\hat{Y}_h)}{\sigma(\hat{Y}_h)} \sim N(0,1), \quad \frac{s^2(\hat{Y}_h)}{\sigma^2(\hat{Y}_h)} \sim \frac{\chi^2_{n-2}}{n-2}$$

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

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

Confidence Interval for $E(Y_h)$

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



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

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 $Y_{h(new)}$

Assume normal error regression model

- $Y_{h(new)} = \beta_0 + \beta_1 X_h + \epsilon_h$
- Assume $\beta_0, \beta_1, \sigma^2$ are known. Then the 1-α prediction limits for $Y_{h(new)}$ is

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

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

Prediction Intervals for $Y_{h(new)}$

estimate

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

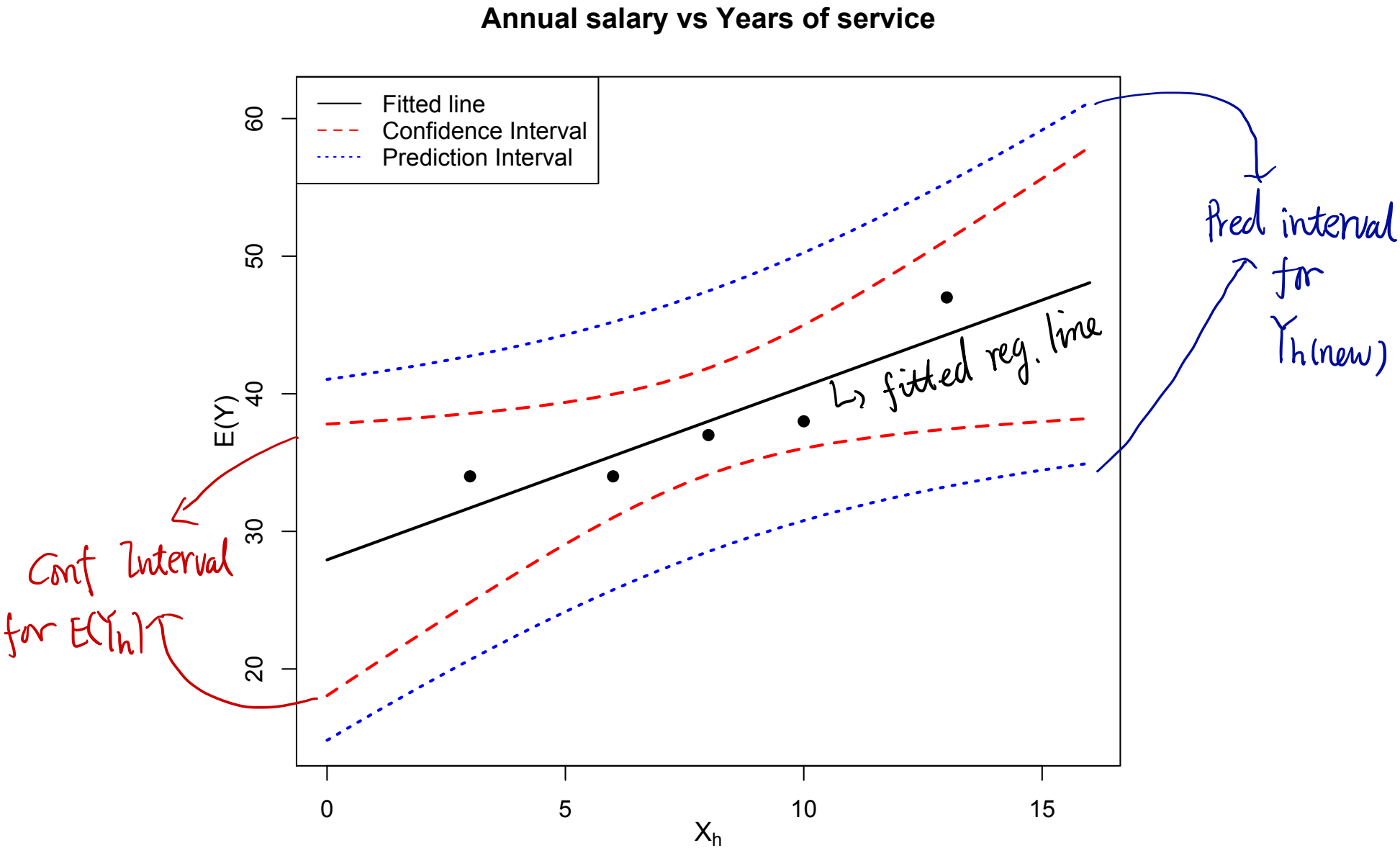
- Replace the σ^2 with MSE, an unbiased estimator of σ^2_{pred} is

$$s^2_{pred} = MSE\{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\}$$

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

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

Example: Confidence intervals vs Prediction intervals



Confidence interval vs Prediction intervals

- 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 α level.

Confidence interval vs Prediction intervals

- Prediction inteval 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 concepturally,
  - A CI represents an inference on a parameter and is an interval that is intended to cover the avalue of the parameter.
  - A PI, on the other hand, is a statement about the value to be taken by a random variable, the new obvervation $Y_{h(new)}$.

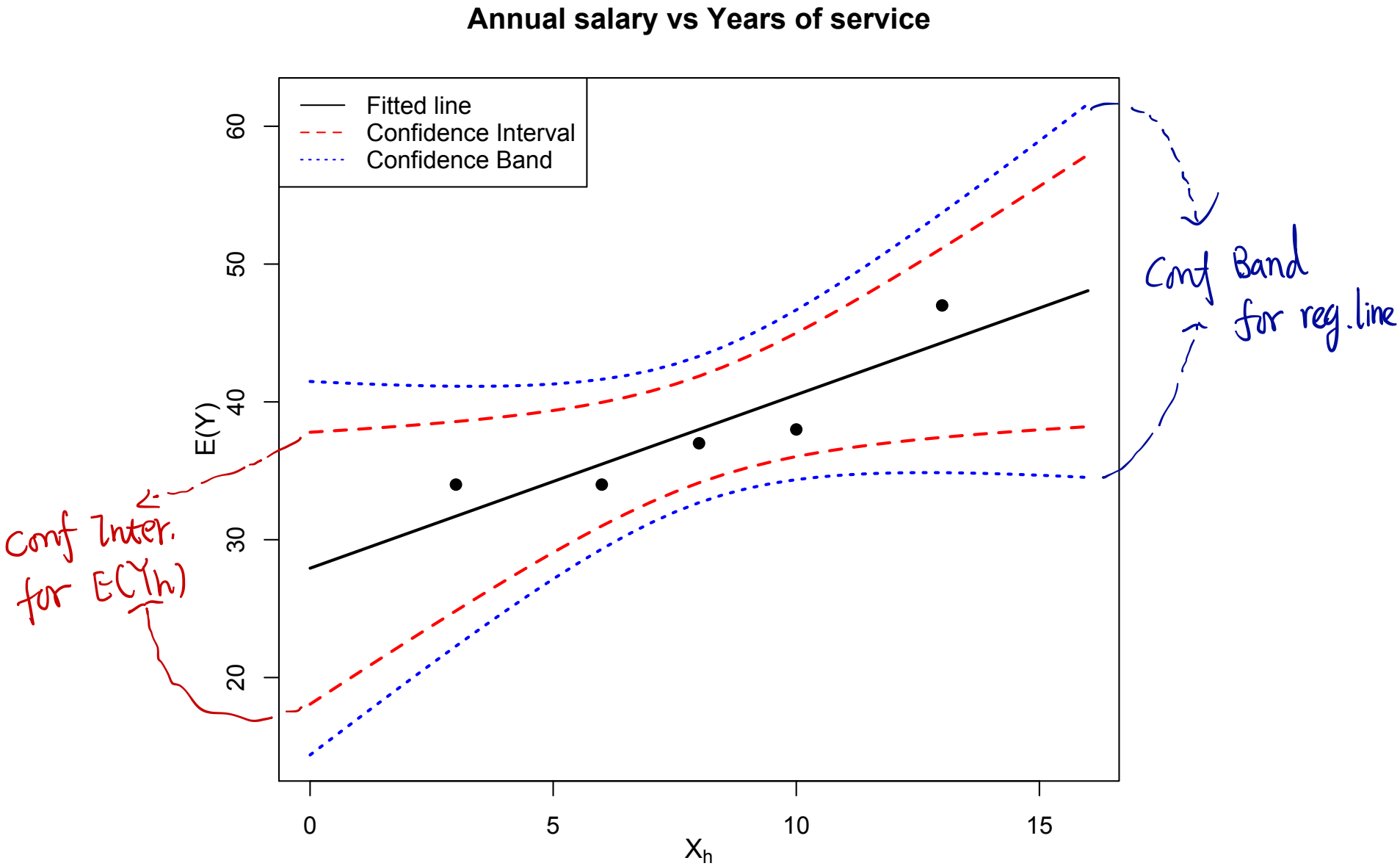
Confidence Band for Regression line

- 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

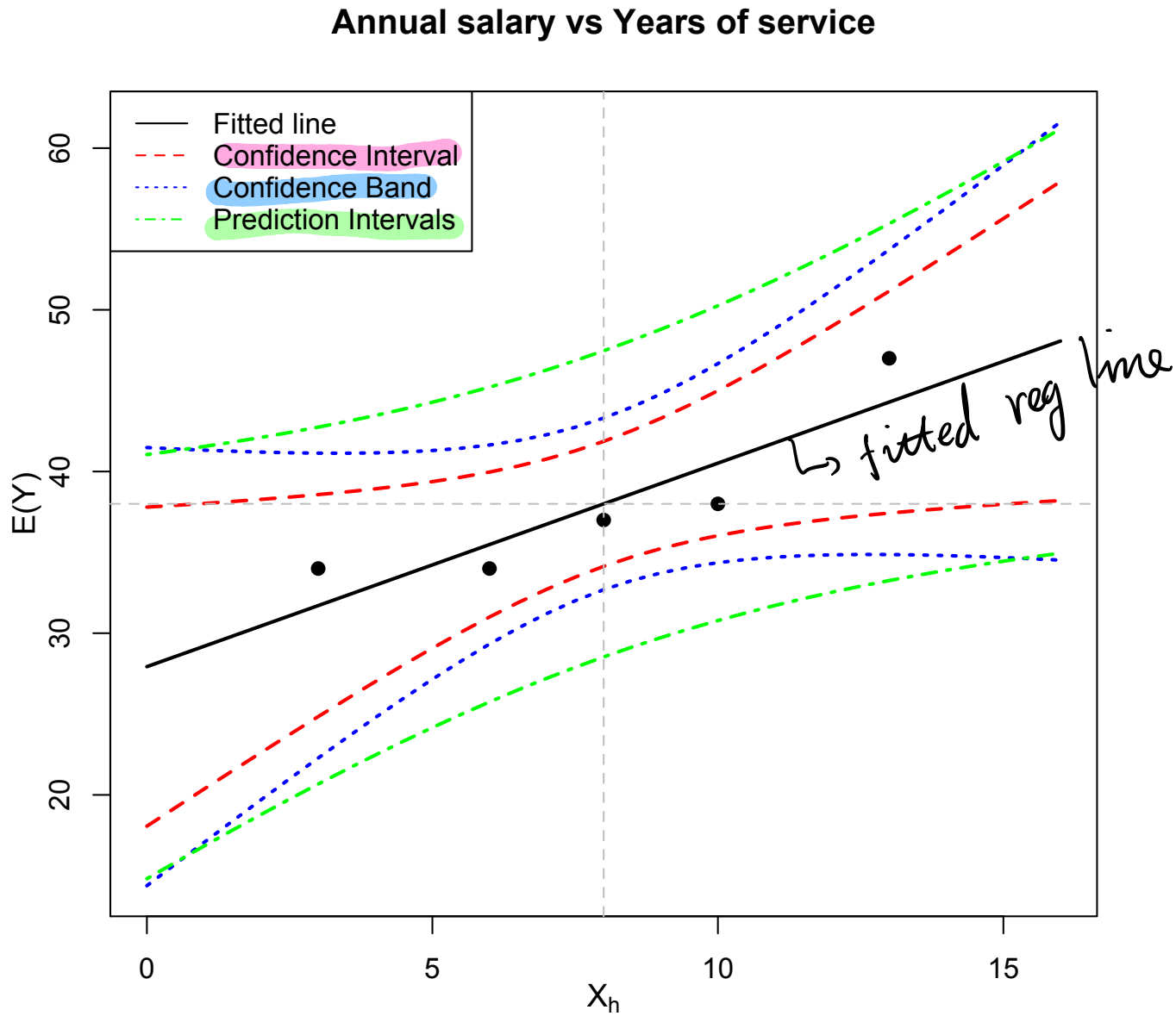
$$\hat{Y}_h \pm W\, s(\hat{Y}_h) = (b_0 + b_1 X_h) \pm W \cdot \sqrt{MSE\{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}\}}$$

  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 substantically wider than the confidence limits for the mean response $E(Y_h)$ at X_h.

Confidence Band vs Confidence Limits



Confidence Band vs Confidence Limits vs Prediction Limits

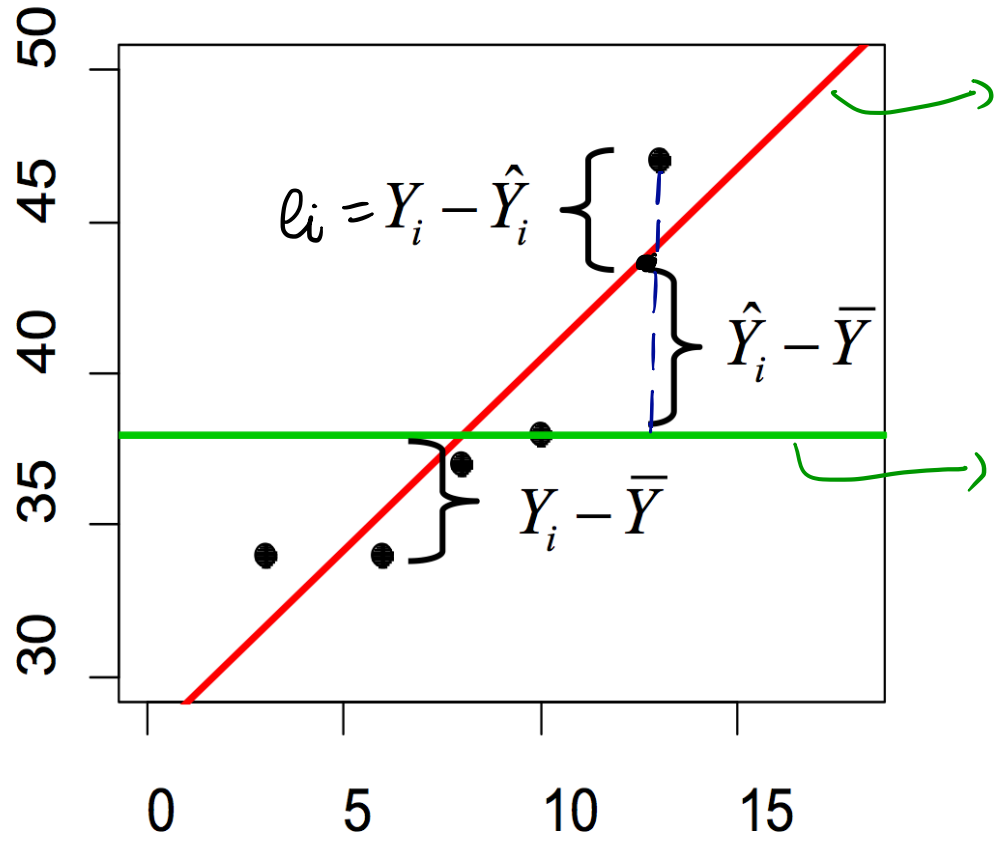


Analysis of Variance (ANOVA) approach to Regression Analysis

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



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

$$\underbrace{\sum_i (Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum_i (Y_i - \hat{Y}_i)^2}_{SSE} + \underbrace{\sum_i (\hat{Y}_i - \bar{Y})^2}_{SSR}$$

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

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$= \sum_{i=1}^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$$

$$= \sum_{i=1}^{n} \{ (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \}^2$$

$$= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

$$= \underbrace{\sum_{i=1}^{n} e_i^2}_{} + \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}_{} + 2 \sum_{i=1}^{n} e_i (\hat{Y}_i - \bar{Y})$$

$$= SSE + SSR + 2 \underbrace{\sum_{i=1}^{n} e_i \hat{Y}_i}_{=0} - 2\bar{Y} \underbrace{\sum_{1}^{n} e_i}_{=0}$$

$$= SSE + SSR$$

Q.E.D.

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

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$\hat{Y}_i = b_0 + b_1 X_i$

$$= \sum_{i=1}^{n} (b_0 + b_1 X_i - \bar{Y})^2$$

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

$$= \sum_{i=1}^{n} (\bar{Y} - b_1 \bar{X} + b_1 X_i - \bar{Y})^2$$

$$= \sum_{i=1}^{n} (b_1 (X_i - \bar{X}))^2$$

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

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 sqaures(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)=σ^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 $\frac{(n-2)MSE}{\sigma^2} \sim \chi_{n-2}^2$

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

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

$$MSR = SSR/1$$

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

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

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

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

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

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

Q.E.D.

for a RV X

$Var(X) = E(X^2) - (EX)^2$

$\Rightarrow E(X^2) = V(X) + (EX)^2$

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

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

$$\frac{MSR}{\sigma^2}|_{H0:\beta_1=0} \sim \chi^2_1, \quad E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}, \quad E(MSE) = \sigma^2$$

MSR > MSE : $\beta_1 \neq 0$

MSR ≈ MSE : $\beta_1 = 0$

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

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

$$\frac{\chi^2_1/1}{\chi^2_{n-2}/(n-2)} \sim F_{1,n-2}$$

- To test H_0
  - Calcuate 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 $H_a: \beta_1 \neq 0$

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

- $$F^* = \frac{b_1^2 \Sigma (X_i - \bar{X})^2}{MSE} = \frac{b_1^2}{MSE / S_{xx}} = \left(\frac{b_1}{\sqrt{MSE / S_{xx}}} \right)^2$$

$$= (t^*)^2 \sim F_{1, n-2}$$

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

ANOVA: F test and t test

- At any given α 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)

```r
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
##                          =MSE^{1/2}
## 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: beta <>0
12*P(t_3>3.53)

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

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

```r
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
```

SSR → MSR = 91.879/1

SSE → MSE = 22.121/3

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

```r
3.53^2
```

```
## [1] 12.4609
```

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 lecture: CH2.8, 2.9, 2.11