STAC67: Regression Analysis

Lecture 5

Sohee Kang

Jan. 27, 2021

2.4 Interval Estimation of $E(Y|X = x_0)$

Suppose that x_0 is a new value of x for which we want to do prediction.

1. Estimation of $\mu_0 = E[Y|X = x_0]$
2. Prediction of Y value for an individual with $X = x_0$

- We use fitted regression model to do both of these
- Estimation of μ_0

$$\mu_0 = \beta_0 + \beta_1 X_0$$

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \hat{Y}_0$$

$$E(\hat{Y}_0) = E(\hat{\beta}_0 + \hat{\beta}_1 X_0) = \beta_0 + \beta_1 X_0$$

$$Var(\hat{Y}_0) = Var(\hat{\beta}_0 + \hat{\beta}_1 X_0) = Var(\hat{\beta}_0) + x_0^2 Var(\hat{\beta}_1) + 2 x_0 Cov(\hat{\beta}_0, \hat{\beta}_1)$$

$$Var(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)\sigma^2 \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$\begin{aligned} Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov(\bar{Y} - \bar{X}\hat{\beta}_1, \hat{\beta}_1) \\ &= Cov(\bar{Y}, \hat{\beta}_1) - \bar{X}\, \underbrace{cov(\hat{\beta}_1, \hat{\beta}_1)}_{Var(\hat{\beta}_1)} \\ &= Cov\left(\frac{1}{n}\sum Y_i, \sum k_i Y_i\right) - \bar{X} \cdot \frac{\sigma^2}{S_{xx}} \\ &= \sum \frac{k_i}{n} Var(Y_i) = \sigma^2 \cdot 0 - \bar{X} \cdot \frac{\sigma^2}{S_{xx}} = -\bar{X}\frac{\sigma^2}{S_{xx}} \end{aligned}$$

Let's derive the variance formula:

Derivation of $Var(\hat{Y}_0)$

Thus $Var(\hat{Y}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{Sxx} \right) \sigma^2 + \frac{x_0^2 \sigma^2}{Sxx} - \frac{2 x_0 \bar{x}}{Sxx}$

$$= \left(\frac{1}{n} + \frac{1}{Sxx} (x_0 - \bar{x})^2 \right) \sigma^2$$

1. as n increases $\sigma^2_{\hat{Y}_0}$ increases
2. as distances between x_0 and $\bar{x}$ increases, $\sigma^2_{Y_0}$ increases
3. as variance of x increases, $\sigma^2_{Y_0}$ decreases

Confidence Interval for $\mu_0 = E(Y|X = x_0)$

- A $(1-\alpha)\%$ confidence interval for μ_0:

$$\hat{Y}_0 \pm t(1-\alpha/2; n-2) SE(\hat{Y}_0)$$

- Exercise: Obtain 95% confidence interval for the mean crime rate for states of high school graduate rate of 80%.

$\hat{Y}_0 = 20517.6 - 170.58(80) \approx 6871.2$

$SE(\hat{Y}_0) = \hat{6}\sqrt{\frac{1}{n} + \frac{1}{S_{xx}}(x_0 - \bar{X})^2} \approx 263.57$

$\hat{6} = \sqrt{MSE} \approx 2356 \approx$ r code "residual standard error"

$6871.2 \pm \underbrace{t(0.975, 82)}_{\text{use r code "qt(0.975,82)"}}; SE(\hat{Y}_0) = 1.989319 \times 263.57 =$

```r
new.data = data.frame(X=80)
predict(fit, new.data, interval="confidence")
```

```
##        fit      lwr      upr
## 1 6871.585 6347.116 7396.054
```

2.5 Prediction of new observation

The value of Y for an individual with $X = x_0$ is:

$$Y_0 = \beta_0 + \beta_1 X + \epsilon_0$$

- Estimate of Y_0: $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0 = \hat{\mu}_0$
- Prediction error: $e_0 = Y_0 - \hat{Y}_0$
- $Var(e_0) = Var(Y_0 - \hat{Y}_0) = Var(Y_0) + Var(\hat{Y}_0)$

$$= \sigma^2 + \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\sigma^2$$

$$= \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\sigma^2$$

- $100 \times (1 - \alpha)\%$ **Prediction Interval** for a new observation $Y_{0(new)}$ with $X = x_0$ is:

$$\hat{Y}_0 \pm t(1 - \alpha/2; n - 2)s_{\{pred\}}$$

Exercise

- Exercise: Obtain 95% prediction interval for the crime rate for states of high school graduate rate of 80%.

$\hat{Y}_0 = 6871.2$ $\quad S_{pred} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}$

$= 2370.7$

Pred Int $= [2154.93, 11587.5]$

```r
new.data = data.frame(X=80)
predict(fit, new.data, interval="prediction")
```

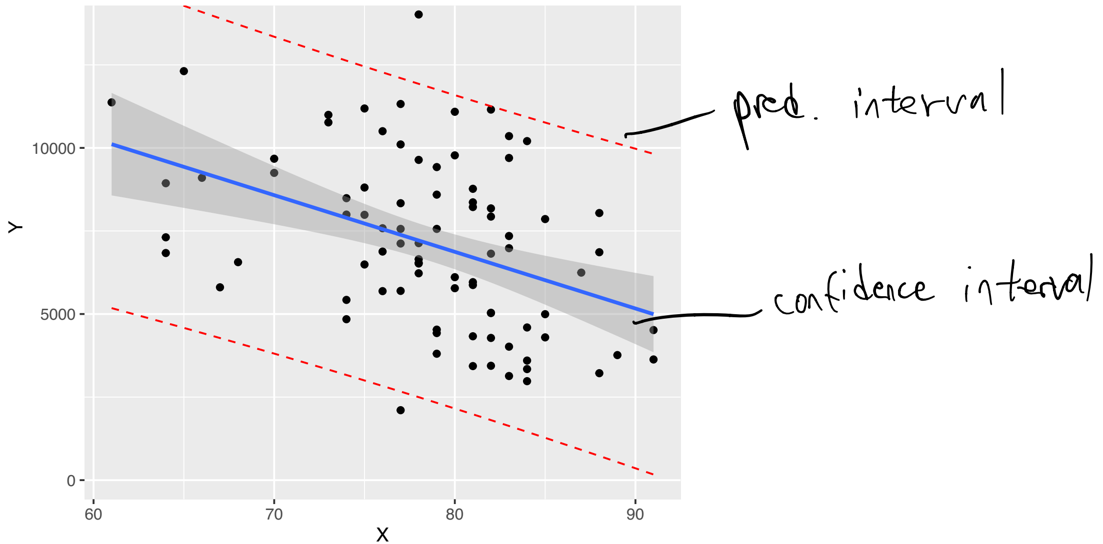
```
##          fit     lwr      upr
## 1 6871.585 2154.92 11588.25
```

Graph Prediction Intervals (using ggplot)

```
Crime.Pred <- predict(fit, interval="prediction")
```

```
## Warning in predict.lm(fit, interval = "prediction"): predictions on current data refer to _future_ responses
```

```
new.df = cbind(Crime,Crime.Pred)
library(ggplot2)
ggplot(data=new.df, aes(X, Y)) +
geom_point()+
geom_line(aes(y=lwr), color="red", linetype="dashed")+
geom_line(aes(y=upr), color="red", linetype="dashed")+
geom_smooth(method=lm, se=TRUE) + theme(plot.margin=unit(c(-1, 8, 7, 4), "cm"))
```



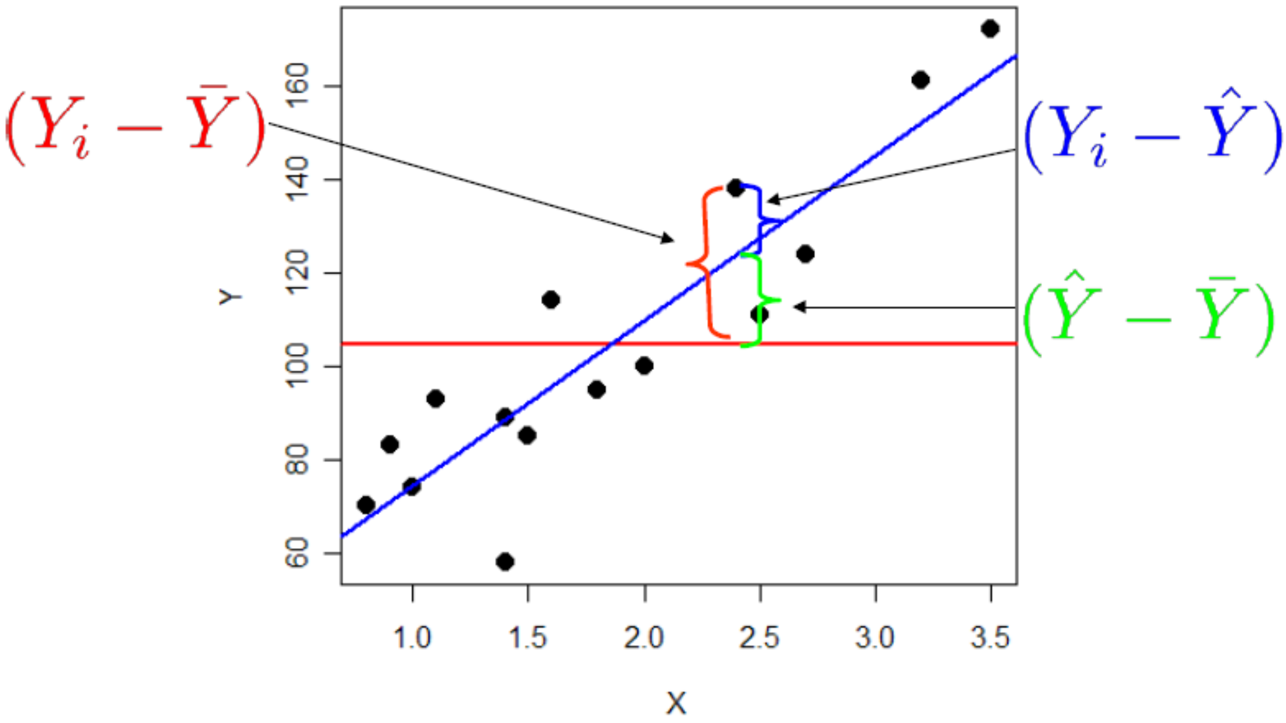
2.7 Analysis of Variance Approach to Regression Analysis

- How well does the least squres fit explain variation in Y?

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

- Total Sum of Squares (SST)
- Model Sum of Squares (SSR): Variation in Y explained by the regression.
- Error Sum of Squares (SSE): Variation in Y that is left explained.

How does that breakdown look on a scatterplot?



Analysis of Variance Table

Source of Variation	Sum of Squares	df	Mean Squares
Regression	SSR		
Error	SSE		
Total	SST		

- The total degrees of freedom is always $n - 1$.
- In the simple regression, the degrees of freedom used by the model is:
- F test for $H_0 : \beta_1 = 0$

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

- Under H_0, $\quad F \sim F(1, n-2)$

F distribution

- The F-distribution with k_1 and k_2 degrees of freedom can be defined as the distribution of the random variable F

$$F = \frac{V_1/k_1}{V_2/k_2},$$

where,

- This is denoted as $F \sim F(k_1, k_2)$
- we can show that under $H_0 : \beta_1 = 0$,

$$\frac{MSR}{MSE} \sim F(1, n-2)$$

Relationship b/w F-test and t-test

- We can rewrite SSR using the regression estimator:

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

$$F^* = \frac{SSR/1}{SSE/(n-2)} =$$

- In the simple regression, this is equivalent to the test of

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

2.9 Descriptive Measure of Linear Association b/w X and Y

$$SST \quad = \quad SSR \qquad + \qquad SSE$$

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

$\dfrac{SSR}{SST}$:

- A "good" model should have a large $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- R^2: Coefficient of determination

Coefficient of Determination

1. $0 \leq R^2 \leq 1$
2. In simple regression, $R^2 = r^2$

- Show that why $\frac{MSR}{MSE} \sim F(1, n-2)$.

$$\frac{\left(\frac{10^3}{10} + 100\right) - \left(\frac{0^3}{10} + 100\right)}{10}$$

Example (Crime Rate)

```r
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df    Sum Sq  Mean Sq F value    Pr(>F)
## X          1  93462942 93462942  16.834 9.571e-05 ***
## Residuals 82 455273165  5552112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Test whether or not there is a linear association between crime rate and percentage of high school graduates using F test. Show the numerical equivalence of two test statistics and decision rules.

2. Compute R^2 and r.