STAC67: Regression Analysis

Lecture 6

Sohee Kang

Jan. 28, 2021

2.7 Analysis of Variance Approach to Regression Analysis

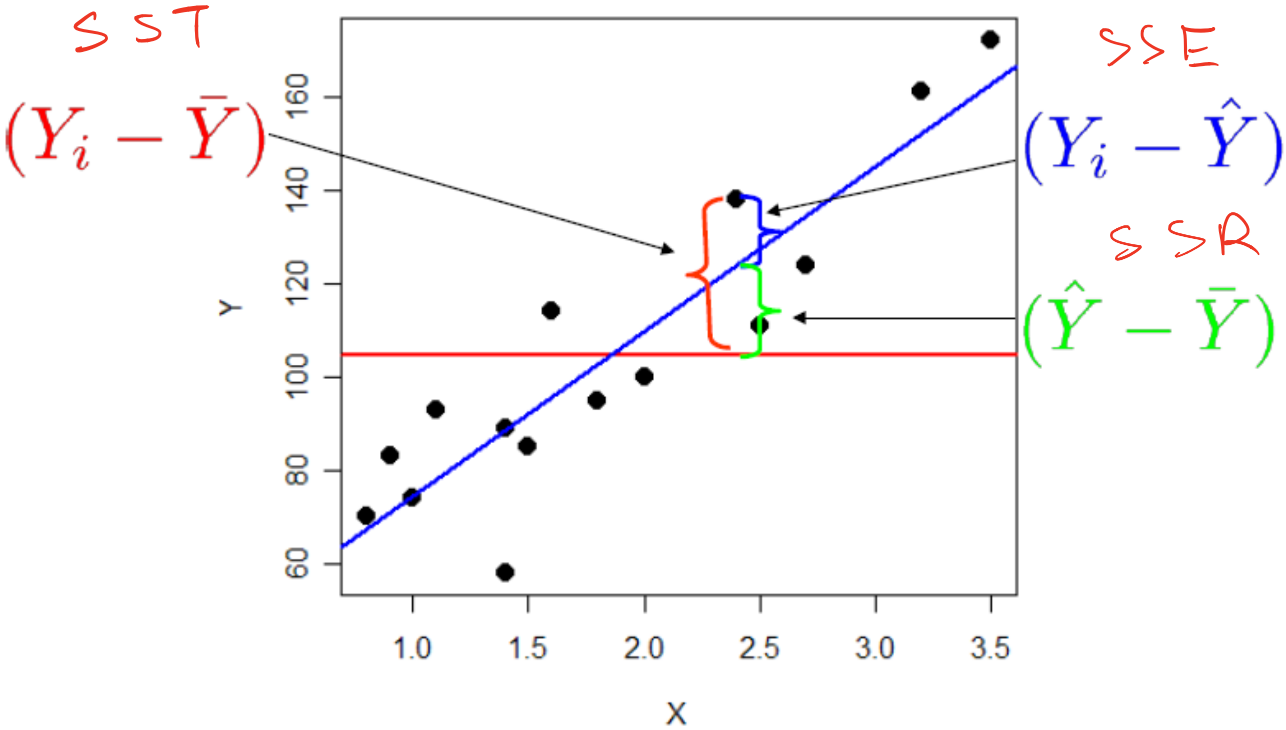
- How well does the least squres fit explain variation in Y?

$$\mathrm{Var}(Y) = \overbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}^{S_{YY} = SST = SSTO} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$$

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

- Total Sum of Squares (SST) : total variation in Y
- 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	1	SSR/1
Error	SSE	n-2	SSE/n-2 = MSE
Total	SST	n-1	

- The total degrees of freedom is always $n - 1$.
- In the simple regression, the degrees of freedom used by the model is: 1
- 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,

$V_1 \sim \chi^2_{k_1}$
$V_2 \sim \chi^2_{k_2}$

V_1, V_2 are independent

- 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:

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \sum(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y})^2$$

$$= \sum(\bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i - \bar{Y})^2 = \sum \hat{\beta}_1^2 (X_i - \bar{X})^2 = \hat{\beta}_1^2 S_{xx}$$

$$F^* = \frac{SSR/1}{SSE/(n-2)} = \frac{\hat{\beta}_1^2 \sum(X_i - \bar{X})^2}{MSE} = \frac{\hat{\beta}_1^2}{\frac{MSE}{S_{xx}}} = \left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}\right)^2 = t^2$$

- 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

$$SST = SSR + SSE$$

Measure of variance in Y without taking covariates into account.

Amount of variability explained by model

Amount of variability left after fitting a linear regression for the covariates

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

$R^2 = \frac{SSR}{SST}$: Proportion of variability of total sum of squares explained by model with predictor X. :

- 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, $R^2 = r^2$ sample corralation coefficient.

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

$$MSR = \left(\frac{\hat{\beta}_1 - \beta}{\frac{\sigma}{S_{xx}}} \right)^2 \sim \chi^2_1$$

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

$$F = \frac{\frac{SSR}{\sigma^2/1}}{\frac{SSE}{\sigma^2/n-2}} = \frac{MSR}{MSE} = F(1, n-2)$$

Example (Crime Rate)

```
anova(fit)
```

n=82+2 = 84

```
## 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
```

SSR MSR

SSE MSE

p < α thus reject

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.

   $H_0: \beta_0 = 0 \quad H_1: \beta_0 \neq 0$

   $F = \frac{MSR}{MSE} = 16.834$ Thus reject the null hypothesis.

2. Compute R^2 and r.

   $R^2 = 0.17$, thus there are more underlying covariates. Eg. Not a good model.

Case Study

- Simple Linear Regression Model Example: COVID-19
- Chicago covid-19 dataset: "Covid-19.csv" (small sized dataset of Chicago with features such as tests vs cases count)

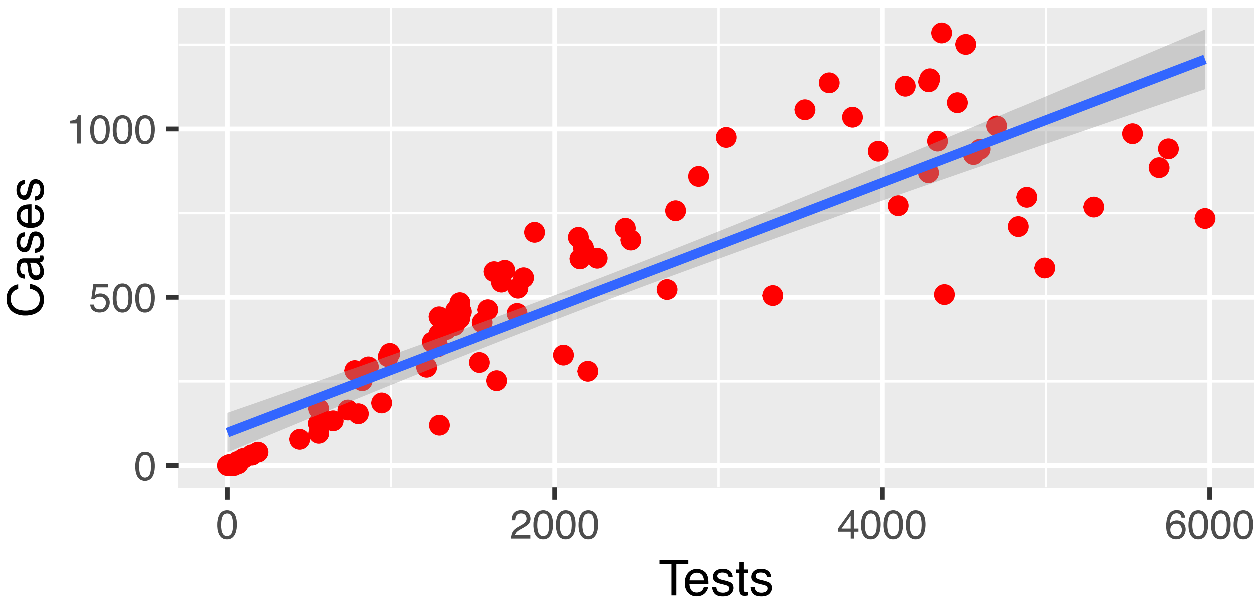
1. Draw a scatter plot of the data.
2. Find the correlation coefficient and test the hypothesis that number of tests and number of cases are lienarly related. What is your conclusion?
3. Do a regression to show how well number of tests can be used to predict the number of cases and find the estimated intercept and slope as well as the estimate of the standard deviation σ.
4. Provide a confidence interval for the slope.
5. Predict the number of cases when the number of tests equals to 10, 100, 1000, 5000.
6. Give 95% confidence intervals for the mean number of cases for given the number of tests and also 95% prediction intervals.

Rcodes

```
Covid = read.csv("COVID-19.csv",header=TRUE)

library(ggplot2)

ggplot(data=Covid, aes(Tests, Cases))+geom_point(col="red") +geom_smooth(method="lm")
```



R codes

```
cor.test(Covid$Tests, Covid$Cases)
```

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2)$$

$n = 87 + 2 = 89$

```
##
##  Pearson's product-moment correlation
##
## data:  Covid$Tests and Covid$Cases
## t = 17.405, df = 87, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8244944 0.9206675
## sample estimates:
##       cor
## 0.8814077
```

```
fit = lm(Cases~Tests, data=Covid)
summary(fit)
```

```
##
## Call:
## lm(formula = Cases ~ Tests, data = Covid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -472.73  -99.45    1.19  114.42  376.91
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 97.77689   29.54996   3.309  0.00136 **
## Tests        0.18572    0.01067  17.405  < 2e-16 ***
```

R codes

```
fit = lm(Cases~Tests, data=Covid)
attach(Covid)
new.data = data.frame(Tests= c(10, 100, 1000, 5000))
predict(fit, new.data, interval="confidence")
```

ANOVA table : anova (fit)

```
##            fit       lwr       upr
## 1   99.63412  41.06609  158.2022
## 2  116.34920  59.25869  173.4397
## 3  283.49999 239.32699  327.6730
## 4 1026.39241 956.00421 1096.7806
```

```
predict(fit, new.data, interval="predict")
```

```
##            fit        lwr       upr
## 1   99.63412 -250.66579  449.9340
## 2  116.34920 -233.70671  466.4051
## 3  283.49999  -64.68252  631.6825
## 4 1026.39241  673.92353 1378.8613
```

```
detach(Covid)
```

Chapter 3. Diagnostics and Remedial Measures

Model assumptions

$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

Key assumptions of our linear regression model:

(L) The conditional mean of Y is linear in X.

(N) Normality of Error terms, $\epsilon \sim N(0, \sigma^2)$

(I) Independent/uncorrelated error terms, $Cov(\epsilon_i, \epsilon_j) = 0, \quad for \quad i \neq j$

(E) Equal (constant) error variance: $Var(\epsilon_i) = \sigma^2$

Inference and prediction relies on this model being true!

- If the model assumptions do not hold, then
  - predition can be systematically biased
  - standard errors and confidence intervals are wrong
- We will focus on using graphical methods to detect the violations of the models.

Checking Assumptions

- Anscombe's quartet: comprises four datasets that have similar statistical properties - even the regression lines and R^2 are the same.

```
attach(anscombe <- read.csv("anscombe.csv"))
c(cor1=cor(x1,y1), cor2=cor(x2,y2), cor3=cor(x3,y3), cor4=cor(x4,y4))
```

```
##      cor1      cor2      cor3      cor4
## 0.8164205 0.8162365 0.8162867 0.8165214
```

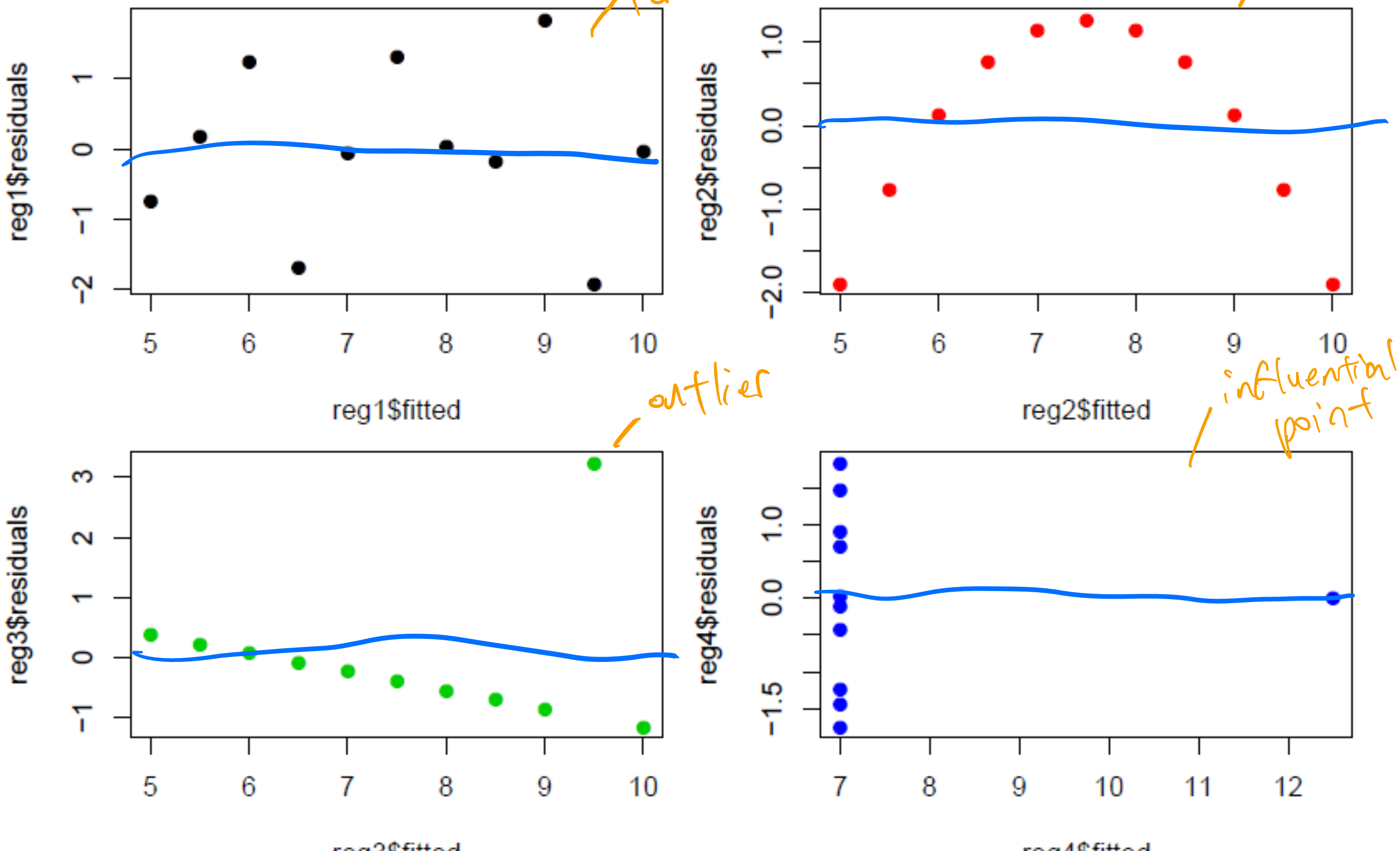
Residual Plots

$\bar{Y}$

- The residuals (plotted against $\hat{Y}$) look totally different.

ideal data

not linear



outlier

influential point