Stat 525 Regression Analysis

Lecture 2 : Inferences in Regression and Correlation Analysis

Zheni Utic, utic@math.umass.edu
Department of Mathematics and Statistics, University of Massachusetts Amherst

Outline

- Simple Linear Regression Model, $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

  · β_0 is the intercept of the line which is the mean of the conditional probability distribution of Y at $X = 0$

  · β_1 is the slope of the line which is the change in the mean of the conditional probability distribution of Y per unit increase in X

  · ϵ_i are uncorrelated random errors ($\sigma^2(\epsilon_i) = \sigma^2$ and $\sigma(\epsilon_i, \epsilon_j) = 0$)

  · Why we need the distribution assumption of ϵ?

  $\rightarrow$ under the normality assumption of ϵ_i (i.e., $\epsilon_i \sim N(0, \sigma^2)$), ϵ_i are independent, and $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ and Y_i are independent

- Topics

  · confidence interval and tests about β_0 and β_1

  · confidence interval about $E(Y)$ for given X

  · prediction interval for a new observation Y

  · ANOVA approach to test about β_1

  · descriptive measures of linear association between variables

2.1 Inferences concerning β_1

- Our interests are

  · point estimator for β_1 and its sampling distribution

  · $100(1-\alpha)\%$ confidence interval for β_1

  · $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

  - $\beta_1 = 0$: there is no linear association between Y and X

- Sampling distribution of b_1 and $\frac{b_1-\beta_1}{s(b_1)}$ (under repeated sampling)

  · different values of b_1 that would be obtained from repeated sampling where the levels of X are the same as in the data set

  · $b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(X_i-\bar{X})^2}\right)$

  · unbiased estimator of $\sigma^2(b_1)$: $s^2(b_1) = \frac{MSE}{\sum(X_i-\bar{X})^2}$

  · $\frac{b_1-\beta_1}{s(b_1)} \sim t_{(n-2)}$ where $s(b_1) = \sqrt{s^2(b_1)}$.

- $100(1-\alpha)\%$ confidence interval for β_1

  $\cdot$ $b_1 \pm t_{1-\alpha/2;n-2} s(b_1)$

  where $1-\alpha = P\left(t_{\alpha/2;n-2} \leq \frac{b_1-\beta_1}{s(b_1)} \leq t_{1-\alpha/2;n-2}\right)$.

- Two -sided test concerning β_1

  $\cdot$ $H_0 : \beta_1 = \beta_{10}$ (e.g., $\beta_{10} = 0$) vs $H_1 : \beta_1 \neq \beta_{10}$

  $\cdot$ test statistic : $t^\star = \frac{b_1-\beta_{10}}{s(b_1)} \sim t_{n-2}$ under H_0

  $\cdot$ decision rule for given α and observed $t^\star$, $t^\star_{obs}$

  i) do not reject H_0 if $|\ t^\star_{obs}\ | \leq t_{1-\alpha/2;n-2}$ or associated p-value $> \alpha$

  ii) reject H_0 if $|\ t^\star_{obs}\ | \geq t_{1-\alpha/2;n-2}$ or or associated p-value $< \alpha$

- One -sided test concerning β_1

  · $H_0 : \beta_1 \leq \beta_{10}$ vs $H_1 : \beta_1 > \beta_{10}$

  · test statistic : $t^\star = \frac{b_1 - \beta_{10}}{s(b_1)} \sim t_{n-2}$ under H_0

  · decision rule for given α and observed $t^\star$, $t^\star_{obs}$

  i) do not reject H_0 if $t^\star_{obs} \leq t_{1-\alpha;n-2}$ or associated p-value $> \alpha$

  ii) reject H_0 if $t^\star_{obs} \geq t_{1-\alpha;n-2}$ or associated p-value $< \alpha$

2.2 Inferences concerning β_0

- Our interests are

  · point estimator for β_0 and its sampling distribution

  · $100(1-\alpha)\%$ confidence interval for β_0

  · $H_0 : \beta_0 = 0$ vs. $H_1 : \beta_0 \neq 0$

  [note] they are valid only if the range of X includes 0

- Sampling distribution of b_0 and $\frac{b_0-\beta_0}{s(b_0)}$ (under repeated sampling)

  · different values of b_0 that would be obtained with repeated sampling when the levels of the X are held constant from sample to sample.

  · $b_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n}+\frac{\bar{X}^2}{\sum(X_i-\bar{X})^2}\right]\right)$

  · estimator of $\sigma^2(b_0)$: $s^2(b_0) = MSE\left[\frac{1}{n}+\frac{\bar{X}^2}{\sum(X_i-\bar{X})^2}\right]$

  · $\frac{b_0-\beta_0}{s(b_0)} \sim t_{n-2}$ where $s(b_0) = \sqrt{s^2(b_0)}$.

- $100(1-\alpha)\%$ confidence interval for β_0

  · $b_0 \pm t_{1-\alpha/2;n-2} s(b_0)$

  where $1-\alpha = P\left(t_{\alpha/2;n-2} \leq \frac{b_0-\beta_0}{s(b_0)} \leq t_{1-\alpha/2;n-2}\right)$.

- Two-sided test concerning β_0

  · $H_0 : \beta_0 = \beta_{00}$ vs $H_1 : \beta_0 \neq \beta_{00}$

  · test statistic : $t^\star = \frac{b_0-\beta_{00}}{s(b_0)} \sim t(n-2)$ under H_0

  · decision rule for given α and observed $t^\star$, $t^\star_{obs}$

  i) do not reject H_0 if $|\, t^\star_{obs} \,| \leq t(1-\alpha/2; n-2)$ or associated p-value $> \alpha$

  ii) reject H_0 if $|\, t^\star_{obs} \,| \geq t(1-\alpha/2; n-2)$ or associated p-value $< \alpha$

Summary of the regression model in R - Copier example

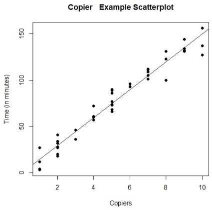
- Let X be the number of copiers serviced and Y be the time spent (in minutes) by the technician

```
   #to upload a data set with "csv" extension in R
>copier=read.csv("C:/Users/stefa/Desktop/STAT 525- Fall 2019/data set/copier.csv",header=TRUE)
>copier                                  #to show the data set in R
       Time   Copiers
1    20       2
2    60       4
3    46       3
................
45   77       5

>reg=lm(Time~Copiers,data=copier)        #to define a linear regression model, where Y is "Time"
                                         and X is number of copiers.
> summary(reg)                           #to call the results of the regression in a table
Call:
lm(formula = Time ~ Copiers, data = copier)

Residuals:
Min        1Q    Median       3Q       Max
-22.7723  -3.7371   0.3334   6.3334  15.4039

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5802     2.8039  -0.207    0.837
Copiers      15.0352     0.4831  31.123   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.914 on 43 degrees of freedom
Multiple R-squared:  0.9575,    Adjusted R-squared:  0.9565
F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

Plot the relationship in R - Copier example

- The estimated equation is $\hat{y} = -0.5802 + 15.0352x$
- We note that the slope $b1 = 15.0352$ implies that for each unit increase in copier quantity, the service time increases by 15.0352 minutes (for quantity values between 1 and 10). If we wish to estimate the time needed for a service call for 5 copiers that would be
- $-0.5802 + 15.0352(5) = 74.5958 minutes$

```
#to plot the relationship between X and Y in R
>plot(copier$Copier,copier$Time,pch=16,xlab="Copiers",ylab="Time (in minutes)",main="Copiers Examp
>abline(reg)              #to add a regression line on the plot
>copier$Copiers           #the variable "Copiers" as a vector from the data set "copier"
>copier$Time              #the variable "Time" as a vector from the data set "copier"
```



Confidence intervals for beta in R - Copier example

- 95% CI for β_1 is
- $15.0352 + t_{1-.025,43}(0.4831) = 16.009486.$
- $15.0352 - t_{1-.025,43}(0.4831) = 14.061010$

```
#to show the confidence intervals in R
>confint(reg,level=0.95)                #to calculate the 95% CI
 2.5 %    97.5 %
(Intercept) -6.234843  5.074529
Copiers     14.061010 16.009486

> confint(reg,level=0.90)               #to calculate the 90% CI
5 %      95 %
(Intercept) -5.29378  4.133467
Copiers     14.22314 15.847352
```

2.3 Considerations on making Inferences concerning β_1 and β_0

- Validity of fitted regression model and, meaning of b_1 and b_0

  · only valid over the span of range of value in our observed data (not outside of those values)

- Effects of departure from normality

  · inferences concerning β_1 and β_0 might hold as long as the probability distribution of Y are not far from normality for finite sample size

- Interpretation of confidence coefficient, $100(1-\alpha)\%$

  · Suppose one take repeated samples where the X observations are kept at the same levels as in the observed sample and a $100(1-\alpha)\%$ confidence interval is obtained for each sample. Then $100(1-\alpha)\%$ of the intervals will enclose the true value of β_1.

2.4 Estimation of $E(Y_h)$ for a given X_h

- Our interests are

  · point estimator for $E(Y_h) = \beta_0 + \beta_1 X_h$ and its sampling distribution

  · $100(1-\alpha)\%$ confidence interval for $E(Y_h)$

  [note] X_h is a value occurring in the sample or other value of the X within the scope of the model

- Point estimator for $E(Y_h)$: $\hat{Y}_h = b_0 + b_1 X_h$

- Sampling distribution of $\hat{Y}_h$ and $\frac{\hat{Y}_h - E(Y_h)}{s(\hat{Y}_h)}$ (under repeated sampling)

  · $\hat{Y}_h \sim N\left(\beta_0 + X_h\beta_1, \sigma^2\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]\right)$

  · estimator for $\sigma^2(\hat{Y}_h)$ is $s^2(\hat{Y}_h) = MSE\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$

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

- $100(1-\alpha)\%$ confidence interval for $E(Y_h)$ and its meaning

  · $\hat{Y}_h \pm t_{1-\alpha/2;n-2} s(\hat{Y}_h)$

  where $1-\alpha = P\left(t_{\alpha/2;n-2} \leq \frac{\hat{Y}_h - E(Y_h)}{s(\hat{Y}_h)} \leq t_{1-\alpha/2;n-2}\right)$.

  · If one takes repeated sampling where the X observations are kept at the same levels as in the observed sample, and obtains a $100(1-\alpha)\%$ confidence interval for each sample, then $100(1-\alpha)\%$ of the intervals will enclose the true value of $E(Y_h) = \beta_0 + \beta_1 X_h$.

Confidence intervals for $E(Y_h)$ for a given X_h in R - Copier

```
#to show the 95% confidence intervals for E(Yh) for each value that belongs to the data set Xh in R
>predict.lm(reg,se.fit=TRUE,copier,interval="confidence",level=0.95)
$fit
       fit       lwr       upr
1   29.49034  25.44468  33.53600
2   59.56084  56.67078  62.45089
3   44.52559  41.14760  47.90357
................................
44  59.56084  56.67078  62.45089
45  74.59608  71.91422  77.27794

$se.fit
1        2        3        4        5        ....
2.006089 1.433068 1.675012 2.006089 2.389533 ....
$df
[1] 43

#to show the 95% confidence intervals for E(Yh) for a specific value Xh=3 in R:
>predict.lm(reg,se.fit=TRUE,newdata=data.frame(Copiers=3),interval="confidence",level=0.95)
$fit
fit     lwr      upr
1 44.52559 41.1476 47.90357

$se.fit
[1] 1.675012
$df
[1] 43
```

- Assume we are interested in an upper 95% confidence limit for the mean time value when the quantity of copiers is 3.
- $44.52559 + t_{1-.025,43}(1.675012) = 47.90375$

2.5 Prediction of $Y_{h(new)}$ for a given X_h

- Our interests are

  · point prediction for $Y_{h(new)}$ when $X = X_h$ (random outcome from the distribution of Y at $X = X_h$: $Y_{h(new)} \sim N(\beta_0 + \beta_1 X_h, \sigma^2)$) and its probability distribution

  · $100(1-\alpha)\%$ prediction interval for $Y_{h(new)}$

  [note] assume that

  i) $Y_{h(new)}$ is independent of Y used in the regression analysis (so, $\sigma(Y_{h(new)}, \hat{Y}_h) = 0$)
  ii) X_h is a value of the X within the scope of the model

  iii) the fitted model for our original data continues to be suitable for $Y_{h(new)}$

- Point prediction of $Y_{h(new)}$ for given X_h is $\hat{Y}_h = b_0 + b_1 X_h$

- Prediction error, $pred = Y_{h(new)} - \hat{Y}_h$

  · variance of prediction error, $\sigma^2(pred) = \sigma^2(Y_{h(new)} - \hat{Y}_h) = \sigma^2 + \sigma^2(\hat{Y}_h)$

- $\frac{Y_h(new) - \hat{Y}_h}{s(pred)} \sim t(n-2)$ where $s^2(pred) = s^2(Y_h(new) - \hat{Y}_h) = MSE + s^2(\hat{Y}_h)$

- $100(1-\alpha)\%$ prediction limit for $Y_{h(new)}$

  · $\hat{Y}_h \pm t_{1-\alpha/2;n-2} s(pred)$

- Prediction interval for $Y_{h(new)}$ sensitive to departure from normality (Chapter 3)

Prediction intervals for $Y_{h(new)}$ for a given X_h in R - Copier

```
#to show the 95% prediction intervals for Yh(new) for each value that belongs to the data set Xh:
> predict.lm(reg,se.fit=TRUE,copier,interval="prediction",level=0.95)
$fit
      fit        lwr       upr
1   29.49034  11.064899  47.91578
2   59.56084  41.354191  77.76748
3   44.52559  26.235146  62.81603

................................
44  59.56084  41.354191  77.76748
45  74.59608  56.421325  92.77084
$se.fit
1        2        3        4        5        ....
2.006089 1.433068 1.675012 2.006089 2.389533 ....
$df
[1] 43

#to show the 95% prediction intervals for Yh(new) for a specific value Xh=7:
>predict.lm(reg,se.fit=TRUE,newdata=data.frame(Copiers=7),interval="prediction",level=0.95)
$fit
      fit      lwr      upr
1 104.6666 86.39922 122.9339
$se.fit
[1] 1.6119
$df
[1] 43
```

- Let us estimate the future service time value when copier quantity is 7 and create a interval around it. The predicted value is:
- $-0.5802 + 15.0352(7) = 104.6666 minutes$
- a 95% upper prediction limit around the predicted value is:
- $104.6666 + t_{1-.025,43}(9.058051) = 122.9339$

2.7 Analysis of Variance approach to regression analysis

- ANOVA (analysis of variance) table

  · partitioning of the total amount of variance in Y

  · which portion of the variance can be accounted for by our model and what portion is just random error

  · capture as much variance in Y by our model as possible

- Partitioning of variation in the observations Y_i

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

(total deviation = deviation of fitted regression value around mean + deviation around fitted regression line)

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

· $\sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$ and $\sum (\hat{Y}_i - \bar{Y})^2 = b_1^2 \sum (X_i - \bar{X})^2$

· $2\sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = 2\sum_{i=1}^n (\hat{Y}_i - \bar{Y})e_i = 2\sum_{i=1}^n \hat{Y}_i e_i - 2\bar{Y}\sum_{i=1}^n e_i = 0$

· SSTO(total sum of squares) : variation(uncertainty) in Y_i, when no account of X is taken

· SSR (regression sum of squares) : variation in Y_i associated with the regression line, $\hat{Y}_i$

· SSE(error sum of squares) : variation in Y_i, when the regression model utilizing X, $\hat{Y}_i$, is employed

· The larger $SSR/SSTO$, the greater is the effect of the regression in accounting for the total variation in the observations Y_i

- Partitioning of the *df*(degrees of freedom) associated with SS(Sum of Squares)

  · *df* : the number of degrees of freedom is the number of independent observations in a sample of data that are available to estimate a parameter of the population from which that sample is drawn

  · (n-1) in SSTO = 1 in SSR + (n-2) in SSE

- Mean Squares (MS=a sum of squares / corresponding *df*)

  · MSR(regression MS) = SSR/1 = SSR and MSE(error MS) = SSE/(n-2)

- ANOVA(Analysis of Variance)

Source of Variation	SS	df	MS
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$	

· $E(MSR) = \sigma^2 + \beta_1^2 \sum_i (X_i - \bar{X})^2$ and $E(MSE) = \sigma^2$. So $E(MSR)/E(MSE) \geq 1$ and $E(MSR) = E(MSE)$ if $\beta_1 = 0$.

- F test of $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ using ANOVA Table

  · test statistic and its sampling distribution : $F^\star = \frac{MSR}{MSE} \sim F_{1,n-2}$ under H_0

  · decision rule for given α

  i) do not reject H_0 if $F^\star_{obs} \leq F_{1-\alpha;1,n-2}$ or associated p-value $> \alpha$

  ii) reject H_0 if $F^\star_{obs} > F_{1-\alpha;1,n-2}$ or associated p-value $< \alpha$

  (hint : large values of $F^\star_{obs}$ ($>>$1) supports H_a and values of $F^\star_{obs}$ near 1 support H_0. so this is an upper-tail test)

- Equivalence of F test and t test in Chapter 2.1

  · under H_0, $F^\star = \frac{MSR}{MSE} = \frac{SSR}{MSE} = \frac{b_1^2 \sum(X_i - \bar{X})^2}{MSE} = \left(\frac{b_1}{s(b_1)}\right)^2 = (t^\star)^2$ where $s(b_1) = \frac{MSE}{\sum(X_i - \bar{X})^2}$

  · $[t_{1-\alpha/2;n-2}]^2 = F_{1-\alpha;1,n-2}$

  · Under the simple linear regression model, F-test for $H_0 : \beta_1 = 0$ is equivalent to t-test for $H_0 : \beta_1 = 0$

ANOVA table in R - Copier

```
#to create the ANOVA table
> anova(reg)
Analysis of Variance Table

Response: Time
Df Sum Sq Mean Sq F value    Pr(>F)
Copiers    1  76960   76960  968.66 < 2.2e-16 ***
Residuals 43   3416      79
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

2.9 Descriptive Measures of Linear Association between X and Y

- Coefficient of Determination, R^2
  - $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$ = proportion of variance accounted for by our model
  - The closer R^2 is to 1, the greater is the degree of linear association

- Misunderstandings of R^2
  - high R^2 indicates useful predictions? Not necessarily
  - high R^2 indicates a good fit of the estimated regression line? Not necessarily
  - R^2 near zero indicates X and Y are unrelated? Not necessarily

- Use both R^2 and a scatter plot of (X, Y)

- Coefficient of Correlation r (when X and Y are both random)

  · $r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$ as an estimator of $\rho = \frac{\sigma(X,Y)}{\sqrt{\sigma^2(X)\sigma^2(Y)}}$

  · $-1 \leq r \leq 1$

  · The closer r is to +1(-1), the greater is the degree of positive(negative) linear association

  · $r = \pm\sqrt{R^2}$ under simple linear regression model (why?)

  · use both r and a scatter plot of (X, Y)

[Remarks]

- When one uses regression analysis for prediction,

  · basic causal conditions in the period ahead should be similar to those in existence during the period on which the regression analysis is based.

  · the prediction in the regression analysis is conditional on X. In practice, however, X often needs to be predicted.

  · the prediction in the regression analysis may be reasonable if X does not fall far beyond the range of the data on which the regression analysis is based.

- When data are obtained from nonexperimental design,

  · $\beta_1 \neq 0 \Rightarrow^?$ a cause-and-effect relation between X and Y?