STAT2008/STAT6038

Overall Significance of the Regression

Assessing the model

- In our fitting we assume the errors have a particular distribution – that is, $\varepsilon \sim N(0,\sigma_\varepsilon^2)$
  - Normal distribution
  - Mean = 0
  - Constant variance = σ_ε^2
    - If σ_ε^2 is small, then small spread of observations around fitted line
    - If σ_ε^2 is large, then observations have wide spread around fitted line
  - Errors are independent :(the errors are independent from the *observed regressor x, and independent of each other)*

Mathematically

$$E(\varepsilon_i \mid x_i) = 0$$

$$E(\varepsilon_i^2 \mid x_i) = \sigma^2$$

$$E(\varepsilon_i \varepsilon_j \mid x_i x_j) = 0$$

$$E(x_i \varepsilon_i) = 0$$

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

Estimating σ^2

- Recall from section 1

choose $\hat{\beta}_0, \hat{\beta}_1$ to minimise

$$S(\hat{\beta}_0,\hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Differentiating

$$A)\quad \frac{\partial S}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$B)\quad \frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

Residuals

Let $e_i = Y_i - b_0 - b_1 X_i$

e_i = sample residuals. They estimate the (unobservable) ε_i's

Equation (A) $0 = \sum_{i=1}^{n} e_i$

Equation (B) $0 = \sum_{i=1}^{n} X_i e_i$

Therefore	Assumptions
$\bar{e} = 0$	$E(\varepsilon_i) = 0$
$\overline{Xe} = 0$	$E(X_i \varepsilon_i) = 0$

Estimating σ^2

The estimator is based on the sample variance of the e_i's.

We use the *residual sum of squares* (also called the *sum of squared errors, SSE*) divided by its appropriate degrees of freedom, $n-2$,

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

This estimator is known as the Mean Squared Error (MSE)

Estimating σ^2- Matrix Notation

7

$$s_e^2 = \frac{e^T e}{n-2} = \frac{(Y-\hat{Y})^T(Y-\hat{Y})}{n-2}$$

$$= \frac{(Y-HY)^T(Y-HY)}{n-2}$$

$$= \frac{Y^T(I-H)^T(I-H)Y}{n-2} = \frac{Y^T(I-H)Y}{n-2}$$

H and $I-H$ are projection matrices (any matrix, A, is called a projection matrix if it satisfies the identities, $A^T = A$ and $AA = A^2 = A$; see Section IV and the first set of tutorial exercises

Assumptions

8

- □ It turns out there are some issues with using our sample residuals to construct the estimator
- □ Leads to an adjustment that ensures that the estimator is unbiased

Leverage (see page 7 of brick)

9

$$Var(e) = \sigma^2(I-H)$$

h_{ij} will usually be non-zero

- □ We can see that if the leverage (the measure of the influence of a data point) is not zero the residuals are not uncorrelated!
- □ The off diagonals are not zero.

Independence

10

- □ Furthermore, it turns out that the correlations between the residuals are always positive
- □ They tend to be closer to each other in value than we might expect, and, more importantly, closer to each other than the corresponding ε_i^2's
- □ This means that the e_i^2's will **underestimate** the ε_i^2's
- □ Biased!

Why n-2? (see page 8 and 9 of the brick)

11

- □ This means that we must divide by a smaller denominator (n-2) to overcome the correlation and make our estimator unbiased!

$$s^2 = \frac{1}{df.}\sum_{i=1}^{n} e_i^2$$

$$\uparrow$$

$$n-2$$

$$\underline{\text{1st Year}}:\ s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Another Problem

12

- □ The variance of our residuals is not constant

$$\text{Var}\left(\varepsilon_i\right) = \sigma^2 = \text{constant}$$

$$\text{Var}\left(e_i\right) = \sigma^2(1-h_{ii})$$

where h_{ii} is the <u>leverage</u> associated with data point i.

Summary

13

GOOD	BAD
$\sum e_i = 0$	e_i not ind.
$\sum X_i e_i = 0$	$\mathrm{Var}\left(e_i\right) \neq \mathrm{Const} = \sigma^2\left(1-h_{ii}\right)$

Standard errors of the parameter estimates and the MSE

14

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)\,\mathrm{var}(X)}}$$

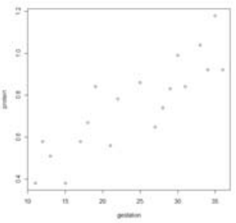
$$s_{b_0} = s_\varepsilon\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\,\mathrm{var}(x)}}$$

S_ε is the square root of the MSE

Example: Protein in Pregnancy

15

- > protpreg<-read.csv("protpreg.csv",header=F)
- > attach(protpreg)
- > protein<-protpreg[,1]
- >gestation<-protpreg[,2]
- > plot(gestation,protein)



Example

16

```
> reg.out<-lsfit(gestation,protein)
> reg.out$coef
 Intercept              X
0.20173770   0.02284426
> reg.diag$std.dev
[1] 0.1150781
> reg.diag$std.err
                    [,1]
Intercept   0.083363149
X             0.003294676
> var(gestation)
[1] 67.77778
> length(gestation)
[1] 19
```

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)\,\mathrm{var}(X)}}$$

$$= \frac{0.1151}{\sqrt{18\times 67.7778}}$$

$$= 0.003295$$

ANOVA

17

- **Analysis of Variance.**
- This part of the output splits the variability in the response into two parts: one part that is variability explained by the model and another part that remains unexplained

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

Total Sum of Squares = Regression Sum of Squares + Error (or Residual) Sum of Squares

SSTO = SSREG + SSE

ANOVA

18

- In the ANOVA (Analysis of Variance) table, the "F" column provides a statistic for testing the hypothesis that $\beta_1 = 0$ against the alternative $\beta_1 \neq 0$
- The test statistic F is the ratio MSR/MSE (the mean square regression term divided by the mean square error term).
- When the MSR term is large relative to the MSE term, then the ratio is large and there is evidence against the null hypothesis.
- For simple linear regression, the statistic F= MSR/MSE has an F distribution with df (1, n - 2).

General Form of ANOVA Table in the Simple Linear Regression Model

Source	d.f.	Sum of Squares	Mean Squares	F Statistics
Regression	1	SSR	MSR=SSR/1	F=MSR/MSE
Error	n-2	SSE	MSE=SSE/(n-2)	
Total	**n-1**	**Variation in Y** **SST**		

ANOVA

Source	D.F.	Sum of Squares	Mean Square	F	P-value
Regression	1	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$\frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{1}$	$\frac{MSREG}{MSE}$	$P(T^2 \geq \frac{MSREG}{MSE})$ $T^2 \sim F(1, n-2)$
Error	$n-2$	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-2}$		
Total	$n-1$	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$			

Significance of the regression

Null and alternative hypothesis:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0.$$

Test Statistic:

$$F = MSR/MSE$$

Decision Rule: Compare F to F distribution on 1 numerator and n-2 denominator degrees of freedom. Reject the null if $F > F_{1,n-2}(1-\alpha)$

Conclusion: If the null is rejected we conclude that there is some relationship between the response and the predictor variable.

lm() function

- lm is used to fit linear models.
- It can be used to carry out regression and analysis of variance
- For instance, a linear regression can be done with the command lm(y ~ x) which means fitting a linear model with y as response and x as predictor.
- lm() allows us to pass dataframes as an argument
- In the next example we are just passing a vector

Example – ANOVA and R

```
> protpreg<-read.csv("protpreg.csv",header=F)
> attach(protpreg)
> protein<-protpreg[,1]
> gestation<-protpreg[,2]
> protpreg.lm<-lm(protein~gestation)
```

ANOVA

```
> anova(protpreg.lm)
Analysis of Variance Table

Response: protein
            Df   Sum Sq    Mean Sq    F value        Pr(>F)
gestation    1   0.63667   0.63667     48.076     2.416e-06 ***
Residuals   17  0.22513   0.01324
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Is this a one sided or two sided test?
- What would it mean to have a small F value?

Conclusion

```
> qf(0.95,1,17)
[1] 4.451322
> qf(0.99,1,17)
[1] 8.39974
```

- We can see that the p-value is much small than 0.05 or 0.01
- Alternatively we can see that the test statistic 48.07606 is much larger than the critical values 4.45 or 8.39

R-Sq – The coefficient of determination

- **Summary of Fit.** R-Squared , gives the proportion of variability in *Y* that is explained by the linear model in *X*. It is a measure of how good the model is: a value of 1 means a perfect fit; a value of 0 suggests a very poor fit (i.e. the linear fit in *X* provides no information about *Y*). The second useful part of this output *s*, the estimate of the error standard deviation σ . NB R-Sq (adj) is used when we consider multiple linear regression

Simple Linear Regression: Determining the Strength and Significance of Association

The strength and significance of the association between the variables of interest is measured by r^2 , or the *coefficient of determination*, which measures the proportion of total variation explained:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}.$$

- Coefficient of Determination measures proportion of total variation of Y explained by the independent variable X
- SSR: variation in Y that is explained by X
- SSE: variation in Y that is unexplained by the variation in X

Sample Correlation Coefficient

The coefficient of determination is equivalent to the square of the sample correlation coefficient,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot SST}}.$$

Hypothesis test for correlation

- H_0: $\rho=0$ (if correlation is zero, then there is no significant linear relationship)
- H_1: $\rho\neq0$
- Test Statistic: $$t = r\sqrt{\frac{n-2}{1-r^2}}$$
- Decision Rule: Compare to a t-distribution with *n*-2 degrees of freedom.
- Conclusion: In terms of whether a significant linear relationship exists.

Example

```
> cor(protein,gestation)
[1] 0.8595157

> cov(cbind(protein,gestation))
             protein gestation
protein   0.04787778  1.548333
gestation 1.54833333 67.777778
```

- The correlation coefficient is 0.85951567
- What is the coefficient of determination?

Hypothesis test in R

31

```
> cor.test(protein,gestation)

        Pearson's product-moment correlation

data:  protein and gestation
t = 6.9337, df = 17, p-value = 2.416e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6648727 0.9448540
sample estimates:
      cor
0.8595157
```

Testing the Slope

32

To test $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$

the test statistic is $T = \dfrac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$

the null hypothesis is rejected at level α if $|T| > t_{(n-2)}(\alpha/2)$.

$$s_{b1} = \frac{s_e}{\sqrt{(n-1)var(x)}}$$

$$= \frac{s_e}{\sqrt{S_{xx}}}$$

The null hypothesis is rejected if p < alpha

Relationship to the F test

33

The regression sum of squares, SSR, can be written as:

$$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}(b_0 + b_1 x_i - \overline{Y})^2$$

$$= \sum_{i=1}^{n}\{(\overline{Y} - b_1\overline{x}) + b_1 x_i - \overline{Y}\}^2 = \sum_{i=1}^{n} b_1^2 (x_i - \overline{x})^2 = b_1^2 S_{xx},$$

which implies that

$$T^2 = \frac{b_1^2 S_{xx}}{s^2} = \frac{SSR/1}{s^2} = \frac{MSR}{s^2} = F.$$

Relationship between T and F

34

```
> qf(0.95,1,17)
[1] 4.451322
> qt(0.975,17)^2
[1] 4.451322
>
```

$$\{t_{n-2}(1-\alpha/2)\}^2 = F_{1,n-2}(1-\alpha)$$

Protein Pregnancy Example

35

```
> summary(protpreg.lm)
Call:
lm(formula = protein ~ gestation)
Residuals:
    Min       1Q          Median          3Q      Max
-0.16853 -0.08720  -0.01009  0.08578  0.20422
Coefficients:
                  Estimate   Std. Error  t value    Pr(>|t|)
(Intercept) 0.201738   0.083363    2.420    0.027 *
gestation   0.022844   0.003295    6.934    2.42e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1151 on 17 degrees of freedom
Multiple R-squared: 0.7388,     Adjusted R-squared: 0.7234
F-statistic: 48.08 on 1 and 17 DF,  p-value: 2.416e-06
```