STA302/1001: Methods of Data Analysis

Instructor: Fang Yao

Chapter 2: Simple Linear Regression (Part II)

Comparing Models

- known as Analysis of Variance (ANOVA)
- a simple example: comparing two regression models

$$\mathrm{E}(Y|X = x) = \beta_0 \text{ v.s. } \mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x$$

- which one to use?
- first model: a horizontal line

  - it says the slope is zero, or

  - cannot help predict Y given X, or

  - X and Y are not related ...

The First Model

- The model is assumed as $\mathrm{E}(Y|X = x) = \beta_0$
- β_0 can be estimated by minimizing $\sum(y_i - \beta_0)^2$, that is, by OLS with only the intercept parameter
- thus $\hat{\beta}_0 = \overline{y}$, the sample mean of $\{y_1, \ldots, y_n\}$.
- residual sum of squares is

$$\sum(y_i - \hat{\beta}_0)^2 = \sum(y_i - \overline{y})^2 = SYY$$

with $n - 1$ degrees of freedom

Which One to Use?

- call $\widehat{\mathrm{E}(Y|X)} = \hat\beta_0$ *fitted model 1*

  call $\widehat{\mathrm{E}(Y|X)} = \hat\beta_0 + \hat\beta_1 x$ *fitted model 2*
- use *fitted model 1* or *fitted model 2*?
- one method is to compare RSS's from two models
- $RSS_1 = SYY$, $RSS_2 = SYY - \frac{(SXY)^2}{SXX}$
- we know $RSS_2 \leq RSS_1$
- the idea is, if adding the slope β_1 does not help much, then RSS_2 should not be much smaller than RSS_1.

Global Min < Local Min

Which One to Use? (cont...)

- key question: how small is small?
- we calculate the difference between RSS_1 and RSS_2, called "sum of squares due to regression" ($SSreg$):

$$\begin{aligned} SSreg &= RSS_1 - RSS_2 \\ &= SYY - \left(SYY - \frac{(SXY)^2}{SXX} \right) \\ &= \frac{(SXY)^2}{SXX} \end{aligned}$$

n obs'n, 1 parameter

n obs'n
2 parameters

- $$\begin{aligned} df \text{ for } SSreg &= df \text{ for } RSS_1 - df \text{ for } RSS_2 \\ &= (n-1) - (n-2) = 1 \end{aligned}$$

due to restriction of the given parameters

The ANOVA Table

- essentially we compare the "standardized version of $SSreg$" v.s. "standardized version of RSS_2"
- we will summarize our comparison in an ANOVA table

Source	df	SS	MSE	F	p-value
Regression	1	$SSreg$	$SSreg/1 \xrightarrow{\text{scale}}$	$MSreg/\hat{\sigma}^2$	get row ① by row ③ − row ②
Residual	$n-2$	RSS	$\hat{\sigma}^2 = RSS/(n-2)$		
Total	$n-1$	SYY			

After fitting linear model with slope

- SS: sum of squares
  MS: mean squares → "MSE" → $\frac{SS}{df}$ (mean square error).

*. If slope helps, RSS_2 should $\ll RSS_1 \Rightarrow SSreg = RSS_1 - RSS_2$ relatively large

we need to standardize by scale.

F-test For Regression

类比：Normal $\xrightarrow{\sigma \to \hat{\sigma}}$ T_{n-2} ; F $\xrightarrow{\sigma \to \hat{\sigma}}$ chi-square

- if the slope β_1 is "useful", then

$$\textcolor{green}{RSS_2} \ll \textcolor{red}{RSS_1} \quad \Rightarrow \quad SSreg \text{ will be relatively large}$$

$$\Rightarrow \quad F = \frac{SSreg/1}{RSS/(n-2)} \text{ will be large}$$

$\hat{\sigma}^2$, estimate of var

this is not true variance $\therefore F \neq$ chi-square

$\dfrac{RSS_1}{\sigma^2} \sim \chi^2_{n-1}$

- F is a rescaled version of $SSreg = \textcolor{red}{RSS_1} - \textcolor{green}{RSS_2}$
- key assumption for F-test: $\textcolor{red}{e_i}$ **are i.i.d.** $\textcolor{red}{N(0, \boxed{\sigma^2})}$, then

$= Var(Y|X)$

if $\beta_1 \neq 0$ 仍为 chi-sq 但不再是 centred at 0

$$\frac{SSreg}{\sigma^2} \sim \chi^2_1 \text{ (if } \beta_1 = 0), \quad \frac{RSS}{\sigma^2} \sim \chi^2_{n-2}, \quad SSreg \perp RSS$$

$\sigma^2 \to$ true var. (unknown)

Orthogonal Decomposition

- recall F-distribution: $F \sim F_{(1,n-2)}$, given $\textcolor{red}{\beta_1 = 0}$

$\to$ one-to-one correspondence to T^2_{n-2}

- what we are doing is a statistical test

$$\textcolor{red}{\text{NH}: \mathrm{E}(Y|X=x) = \beta_0} \quad \text{v.s.} \quad \textcolor{green}{\text{AH}: \mathrm{E}(Y|X=x) = \beta_0 + \beta_1 x}$$

F-dist: 2 indpt r.v. $F(a,b) = \dfrac{RSS_a/a}{RSS_b/b}$, a & b are df of 2 r.v.'s.

F-test For Regression (cont...)

- we compare "the observed value of F" calculated from the sample to the critical value, $F_{(\alpha,1,n-2)}$, the upper-α quantile or $100(1-\alpha)$th percentile of $F_{(1,n-2)}$
- if $F_{obs} > F_{(\alpha,1,n-2)}$, reject NH, use model 2.
- if $F_{obs} \leq F_{(\alpha,1,n-2)}$, don't reject NH (don't say accept)
- Forbe's data, use R function qf(0.95, 1, 15) to find $F_{0.05,1,15} = 4.543$

Source	df	SS	MS	F	p-value
Regression on $Temp$	1	425.639	425.639	2962.79	≈0
Residual	15	2.155	0.144		

- conclusion? # If NH is true, test statistic $F \geq F_{obs}$ is not likely to happen
  $\rightarrow$ p-value $= P(F \geq F_{obs} \mid \beta_1 = 0) \approx 0$
  (i.e. test stat is not as extreme as in observation)

p-value $\neq P(\beta_1 = 0)$, which is P(NH is true)

p-value and Interpretation

- What does it mean? Assuming the NH is true, the probability that the test statistic is at least as extreme as was observed in the sample, e.g., in the previous F-test, p-value$= P(F \geq F_{obs} | \beta_1 = 0) \approx 0$
- a measure of the strength of the evidence against NH in favor of AH, not the probability that NH is true

intepreta-tion of p-value

- compare p-value with significance level α, say $\alpha = 0.05$
- statistical significance v.s. scientific significance
- latter needs the former to confirm

P.I. → ~~$P_{NH} \downarrow$, reject NH~~ 无关联

2 types of errors

Coefficient of Determination, R^2

- definition
$$R^2 = \frac{SSreg}{SYY}$$
Tells you how "useful" your slope is.
- scale-free one number summary
- measure the strength of the relationship between x_i and y_i
- to see this, notice that
- SYY: variability in the data
- $SSreg$: variability in the data explained by the slope

Coefficient of Determination, R^2 (cont...)

- Forbes' data

$$R^2 = \frac{425.63910}{427.79402} = 0.995$$

- it means that the straight line model explains 99.5% of the variability in the data
- another way to look at R^2:

$$R^2 = \frac{SSreg}{SYY} = \frac{(SXY)^2}{SXX\ SYY} = r_{xy}^2$$

- the square of sample correlation between X and Y

Confidence Intervals and Tests

- for "simple problems", if $\hat{\theta}$ is an estimate for θ, then a $100(1-\alpha)\%$ confidence interval (C.I.) for θ is

$$(\hat{\theta} - t_{(\frac{\alpha}{2},d)}\, se(\hat{\theta}), \quad \hat{\theta} + t_{(\frac{\alpha}{2},d)}\, se(\hat{\theta}))$$

  where $se(\hat{\theta})$ is the standard error for $\hat{\theta}$, and $t_{(\frac{\alpha}{2},d)}$ is the value that cuts off $\frac{\alpha}{2} \cdot 100\%$ in the upper tail of the t-distribution with df$= d$

- when to use t-distribution or normal?

- what is the correct way to interpret “a 95% C.I. for θ is $(3.5, 5.6)$?

Confidence Intervals and Tests for β_0

- key assumption: e_i's are i.i.d. $N(0, \sigma^2)$
- for the intercept β_0 the C.I. is

$$(\hat{\beta}_0 - t_{(\frac{\alpha}{2}, n-2)}\, se(\hat{\beta}_0), \quad \hat{\beta}_0 + t_{(\frac{\alpha}{2}, n-2)}\, se(\hat{\beta}_0))$$

where $se(\hat{\beta}_0) = \hat{\sigma}(\frac{1}{n} + \frac{\bar{x}^2}{SXX})^{\frac{1}{2}}$ instead of Z-test

- Hypothesis test: for a pre-fixed β_0^*, say $\beta_0^* = 0$
NH: $\beta_0 = \beta_0^*$, β_1 arbitrary
AH: $\beta_0 \neq \beta_0^*$, β_1 arbitrary
- t-statistic $t = \frac{\hat{\beta}_0 - \beta_0^*}{se(\hat{\beta}_0)}$ and compare to $t_{(\frac{\alpha}{2}, n-2)}$

Confidence Intervals and Tests for β_1

- for the slope β_1

$$\text{C.I.} : \hat{\beta}_1 \pm t_{(\frac{\alpha}{2},n-2)}\, se(\hat{\beta}_1)$$

$$= \hat{\beta}_1 \pm t_{(\frac{\alpha}{2},n-2)} \frac{\hat{\sigma}}{\sqrt{SXX}}$$

- Hypothesis test: similar to β_0
- a special case of NH: $\beta_1 = 0$ v.s. AH: $\beta_1 \neq 0$
- same as comparing "$y = \beta_0$" and "$y = \beta_0 + \beta_1 x$"

Confidence Intervals and Tests – t and F

- doing the t-test

  NH: $\beta_1 = 0$ vs AH: $\beta_1 \neq 0$

  is the same as comparing "$y = \beta_0$" and "$y = \beta_0 + \beta_1 x$" with an F-test
- t-statistic: $t = \frac{\hat\beta_1 - 0}{se(\hat\beta_1)} = \frac{\hat\beta_1}{\hat\sigma/\sqrt{SXX}}$

  $\frac{SSreg}{\hat\sigma^2}$
- $t^2 = \frac{\hat\beta_1^2}{\hat\sigma^2/SXX} = \frac{\hat\beta_1^2 SXX}{\hat\sigma^2} =$ F-statistic from ANOVA Table
- that is, there is a one-to-one correspondence here
- from the fact that the square of t_d is $F_{(1,d)}$
- (then why do we study both the t and the F tests?)

  F is more like a global test
  t is like for single parameter.

They do have the same expected value, but they have different variance

Prediction and Fitted Values

estimation

- first, a simple question
- if $X_1, X_2, \cdots, X_m \sim$ i.i.d. $N(\mu, \sigma^2)$, what is $\text{Var}(\bar{X})$?
- should it be smaller or larger than $\text{Var}(X_i)$?
- prediction: predict the value of y given a new value of x
- denote the new values: x_*, y_*
- x_* is known but y_* is not
- e.g., "income" $= 10 + 20\times$ "year of education"
- You have done 16 years of education. How much are you expected to earn? have nothing to do with the formula above.

one individual tends to have larger variation but a population tends to have a smaller one.

So 10+20×16 is a prediction.
But Maybe 400 is your fitted value.

Prediction

- $\tilde{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$
- $x_* = 16$, $\tilde{y}_* = 100 + 200 \times 16 = 3300$
- You are expected to earn \$3300 a month
- $\tilde{y}_*$ predicts unbiasedly the unobserved y_* (verify)

$$\begin{aligned}\mathrm{Var}(\tilde{y}_* - y_* | \mathbb{X}, x_*) &= \mathrm{Var}(y_*|x_*) + \mathrm{Var}(\tilde{y}_*|\mathbb{X}, x_*) \\ &= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}\right)\end{aligned}$$

$\Downarrow$ The notation on book is incorrect.

$$\mathrm{sepred}(\tilde{y}_* - y_*|\mathbb{X}, x_*) = \hat{\sigma}\left(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}\right)^{\frac{1}{2}}$$

- we can construct a prediction interval for y_*:

$$\tilde{y}_* \pm t_{(\frac{\alpha}{2}, n-2)}\, \mathrm{sepred}(\tilde{y}_*|\mathbb{X}, x_*)$$

Fitted Values

- same "income - years of education" example
- what is the average income of <u>all</u> people who have done 16 years of education?
- this is an estimation problem, not prediction
- estimated by the fitted value

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{with } x = 13$$

- its standard error is $\mathrm{sefit}(\hat{y}|\mathbb{X}, x) = \hat{\sigma}(\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX})^{\frac{1}{2}}$
- compare $\mathrm{sefit}(\hat{y}|\mathbb{X}, x)$ with $\mathrm{sepred}(\tilde{y}_*|\mathbb{X}, x_*)$
- notation in text is a bit confusing

only effective for x_*

Fitted Values (cont...)

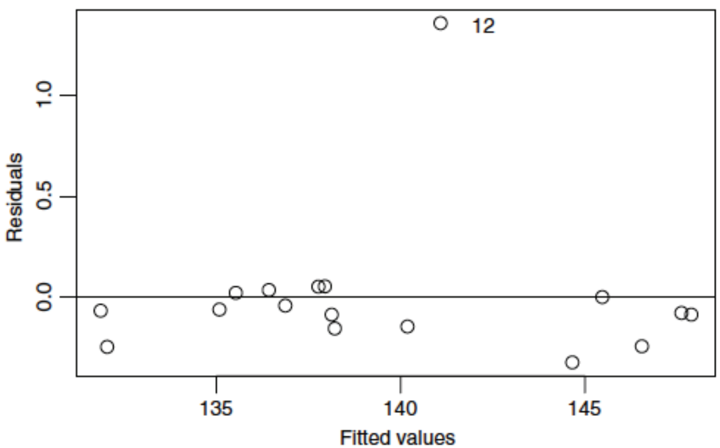
- confidence interval:

$$(\hat{\beta}_0 + \hat{\beta}_1 x) \pm \text{sefit}(\hat{y}|\mathbb{X}, x)[2F(\alpha; 2, n-2)]$$

- note: we are using a F-distribution, not t
- why? another course will tell you...

The Residuals

- definition: $\hat{e}_i = y_i - \hat{y}_i$
- plots can show problems in our modeling
- a useful plot: residuals v.s. fitted values
- Forbes' data



The Residuals (cont...)

- Case 12: possible outlier
- remove Case 12 and re-do the regression
- Summary Statistics for Forbes' Data with All Data and with Case 12 deleted

Quantity	All Data	Delete Case 12
$\hat{\beta}_0$	-42.138	-41.308
$\hat{\beta}_1$	0.895	0.891
se($\hat{\beta}_0$)	3.340	1.001
se($\hat{\beta}_1$)	0.016	0.005
$\hat{\sigma}$	0.379	0.113
R^2	0.995	1.000

A "Good" Residual Plot from Heights Data