1

# STAT2008/STAT6038

Estimation and Prediction using the SLR model

## Why do we need both F and T?

2

- Assuming we have an appropriate model, we then have both the F-test and the T-test to test for the significance of the regression.
- Why don't we just stick with the one distribution – F, since it is the easiest to apply to both multiple regression and simple linear regression models?
- It turns out that we can use T-tests to calculate other useful intervals (including for predictions)

## Recall the Test Statistic for the Slope

3

$$\frac{(b_1-\beta_1)}{s_\epsilon/\sqrt{S_{xx}}} \sim t_{n-2}.$$

$$\frac{(b_1-\beta_1)}{s_{b_1}} \sim t_{n-2}.$$

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

$$= \frac{s_\epsilon}{\sqrt{S_{xx}}}$$

## The Test Statistic for the Intercept

4

$$\frac{(b_0-\beta_0)}{s_\epsilon\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}.$$

$$\frac{(b_0-\beta_0)}{s_{b_0}} \sim t_{n-2}.$$

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

## Testing the intercept

5

- Hypotheses: $H_0$: $\beta_0$=constant
  $H_A$: $\beta_0 \neq$ constant
- Test Statistic: $t = \dfrac{b_0 - \text{constant}}{s_{b_0}}$

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

- Decision Rule: Compare to a t-distribution with n-2 degrees of freedom
- Conclusion: In terms of whether evidence is sufficient to reject null hypothesis.

## For example

6

- For example, we can test if intercept = 0 (Note if intercept = 0, then when x=0, y=0).
- Should always be careful with inferences about the intercept as it may be outside the prediction range, and so will not necessarily have an sensible interpretation.
- For some examples (e.g. calibration), intercept can be very important (equal to a specific value or equal to 0).

## Confidence Interval for the intercept

7

a $100(1-\alpha)\%$ confidence interval for $\beta_0$ can be calculated as:

$$b_0 \pm t_{n-2}(1-\alpha/2)s_e(b_0) = \left(b_0 - t_{n-2}(1-\alpha/2)s_e\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}, b_0 + t_{n-2}(1-\alpha/2)s_e\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}\right).$$

## Special case of regression through the origin

8

- If zero is contained in the above interval, then we might be willing to assume that the true regression line goes through the origin
- But does the range of the data include the origin (or at least nearly include)?

**Or**

- Does it make sense for the line to go through the origin?

## If the intercept does equal 0

9

- If either the data or scientific justification leads us to conclude the intercept is zero then:

$$Y = \beta_1 x + \epsilon,$$

with $\epsilon$ again a mean-zero random variable with variance $\sigma^2$.

## Least Squares Slope Estimator

10

We can then repeat our entire discussion, with the modification that we will start our least-squares estimation from the distance function:

$$d^*(b_1) = d(0, b_1) = \sum_{i=1}^{n}(Y_i - b_1 x_i)^2.$$

Taking derivatives shows that the least-squares estimator of the slope is now

$$b_1 = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}.$$

## Variance of Least Squares Slope Estimator

11

Under the assumption that the model is correct (i.e., that the true regression line does indeed pass through the origin), $b_1$ is still unbiased.

Assuming homoscedasticity (constant)and uncorrelatedness(independence) of the errors, the variance of this estimator is:

$$Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}.$$

## Estimate of the variance

12

An estimate of this quantity can be calculated by substituting $s_e^2$ for $\sigma^2$, where the scale estimate $s_e^2$ is still the $MSE$ of the model, however, this is now defined by:

$$s_e^2 = MSE = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - b_1 x_i)^2.$$

Notice that this estimator now has $n-1$ degrees of freedom, since we no longer need to estimate an intercept term.

## ANOVA

13

Unfortunately, the usual partitioning of the total sum of squares $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ no longer holds.

However a similar breakdown of the quantity

$\sum_{i=1}^{n} Y_i^2$ into

$(SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2) + (\sum_{i=1} \hat{Y}_i^2 = b_1^2 \sum_{i=1}^{n} x_i^2)$ does hold

## Testing significance of the slope

14

we can test the null hypothesis, $H_0 : \beta_1 = 0$, using the $F$-statistic

$$F = \frac{b_1^2 \sum_{i=1}^{n} x_i^2}{s_e^2} \sim F_{1,n-1},$$

and we can construct confidence intervals using the test statistic (AKA pivot):

$$T = \frac{b_1 \sqrt{\sum_{i=1}^{n} x_i^2}}{s_e} \sim t_{n-1},$$

## Confidence Interval

15

So in the special case when the intercept = 0, we can still do tests and make inferences about the slope:

a $100(1-\alpha)\%$ confidence interval is given by:

$$\left(b_1 - t_{n-1}(1-\alpha/2)\frac{s_e}{\sqrt{\sum_{i=1}^{n} x_i^2}}, b_1 + t_{n-1}(1-\alpha/2)\frac{s_e}{\sqrt{\sum_{i=1}^{n} x_i^2}}\right).$$

Note that it is still the case that $T^2 = F$.

## Interval Estimates for new values of X

16

- Often we are interested in predicting the response value for future observations at particular values of the predictor variable.
- Two distinct types of predictions which are of interest.
  - Prediction (for a single future response)
  - Confidence (prediction for the mean or expected response value for observations at a particular predictor value)

## Confidence and Prediction Intervals

17

- A *confidence interval* for the expected value of y (an interval within which we expect to find the average response)
- A *prediction interval* for a single observation of y (an interval within which we expect single observations of the response)

## Estimate of the standard error of prediction

18

$$s\{\hat{Y}(x_g)\} = s\sqrt{\frac{1}{n} + \frac{(x_g - \overline{x})^2}{S_{xx}}}.$$

Assuming normal errors, we can write a $100(1-\alpha)\%$ confidence interval for the expected response $E(Y|x_g)$ as:

$$\hat{Y}(x_g) \pm t_{n-2}(1-\alpha/2)s\{\hat{Y}(x_g)\}.$$

## A *confidence interval* for the expected value of y

19

- 100(1-α)% interval for a given value of x, $x_g$:

$$\hat{y} \pm t_{\alpha/2,n-2} \times s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

- Note that the further away from the x average we are predicting, the wider our prediction interval will be.

## A *prediction interval* for a single observation of y

20

- 100(1-α)% interval for a given value of x, $x_g$:

$$\hat{y} \pm t_{\alpha/2,n-2} \times s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

- Note that the further away from the x average we are predicting, the wider our prediction interval will be.
- Same as formula for prediction interval, but now includes "1+" under the square root: Result – PI is wider than CI for same value of *x*

## Interpretation

21

- A confidence Interval for E(Y) estimates the mean value of y for a population of individuals who all have the same particular value of x.
- The confidence interval includes only the uncertainty in estimating the coefficients, whereas the prediction interval includes *both the uncertainty in estimating the curve coefficients and the uncertainty in measurement.*
- In short – there is less error in estimating a mean value as opposed to predicting an individual value

## CI vs PI

22

- Explain whether a confidence interval for the mean or a prediction interval for the value of y should be used.
  - Estimate the first year ANU average mark for a student with a score of 90 in year 12
  - Estimate the mean first year ANU average mark for students whose scores were 90 in year 12

## Example (CI and PI)

23

**Previously we saw that for the protein in pregnancy data, the least-squares parameter estimates were $b_0 = 0.2017377$ and $b_1 = 0.02284426$, and the estimate of the regression scale was $s_\varepsilon = 0.1150781$. In addition, we can calculate $\bar{x} = 24$ and $S_{xx} = 1220$ (see tutorial one).**

## Back to the Protein in Pregnancy example

24

```
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
```

## First find predicted value

25

Suppose we want to predict the expected protein level of a pregnant woman who has just come to full term, which is typically 38 weeks. Our estimate of the expected protein concentration for such women would be $\hat{Y}(38) = \hat{b}_0 + 38\hat{b}_1 = 1.06982$, and the standard error is

$$s\{\hat{Y}(38)\} = s_e\sqrt{\frac{1}{n} + \frac{(38-\bar{x})^2}{S_{xx}}} = 0.1159781\sqrt{\frac{1}{19} + \frac{(38-24)^2}{1220}} = 0.05314656.$$

## Calculate Confidence Interval

26

Thus, a 95% confidence interval for $E(Y|x_y = 38)$ would be:

$$1.06982 \pm 0.05314656t_{17}(0.975) = (0.9576902, 1.181949)$$

since $t_{17}(0.975) = 2.109816$.

## R

27

```
>gestation<-38
>pr.pred<-predict(protpreg.lm,as.data.frame(gestation),se.fit=T)
> pr.pred$fit
       1
1.06982

> pr.pred$se.fit
[1] 0.05314656

> qt(0.975,17)
[1] 2.109816
```

## Calculate Prediction Interval

28

Now, suppose that we have just taken a protein sample from a pregnant woman who is just about to give birth (i.e., she is at 38 weeks gestation), what would we guess for the value of her protein level? Here, we are asked for a prediction interval and not a confidence interval. So, a 95% prediction interval in this case would be:

$$\hat{Y}(x_y) \pm t_{n-2}(1-\alpha/2)s_e\sqrt{1 + \frac{1}{n} + \frac{(x_y-\bar{x})^2}{S_{xx}}} = \hat{Y}(x_y) \pm t_{n-2}(1-\alpha/2)\sqrt{s_e^2 + [s\{\hat{Y}(x_y)\}]^2}$$

$$1.06982 \pm 2.109816\sqrt{(0.1159781)^2 + (0.05314656)^2}$$

$$= 1.06982 \pm 0.2674354$$

$$\Longrightarrow \quad (0.8023846, 1.337255).$$

## Interval comparison

29

- Which is wider?
- Why?