STOR 455
STATISTICAL METHODS I

Jan Hannig

Exam 1

- Held on Tuesday – regular place and time
- Closed book closed notes, no computers
- Bring your scantron and pencil!

Normal, chi square and t tables will be provided.

- I will have office hours at 12noon on Tuesday. (Provided my red eye lands on time.)

Example: Breast Cancer

- What's the relationship between mean annual temperature and the mortality rate for a type of breast cancer in women? The subjects from regions of Great Britain, Norway, and Sweden.
- Mortality: Mortality index for neoplasms of the female breast
- Temperature: Mean annual temperature (in degrees F)
- The Data (http://www.ncsec.org/cadre2/team6_2/modelII.pdf)

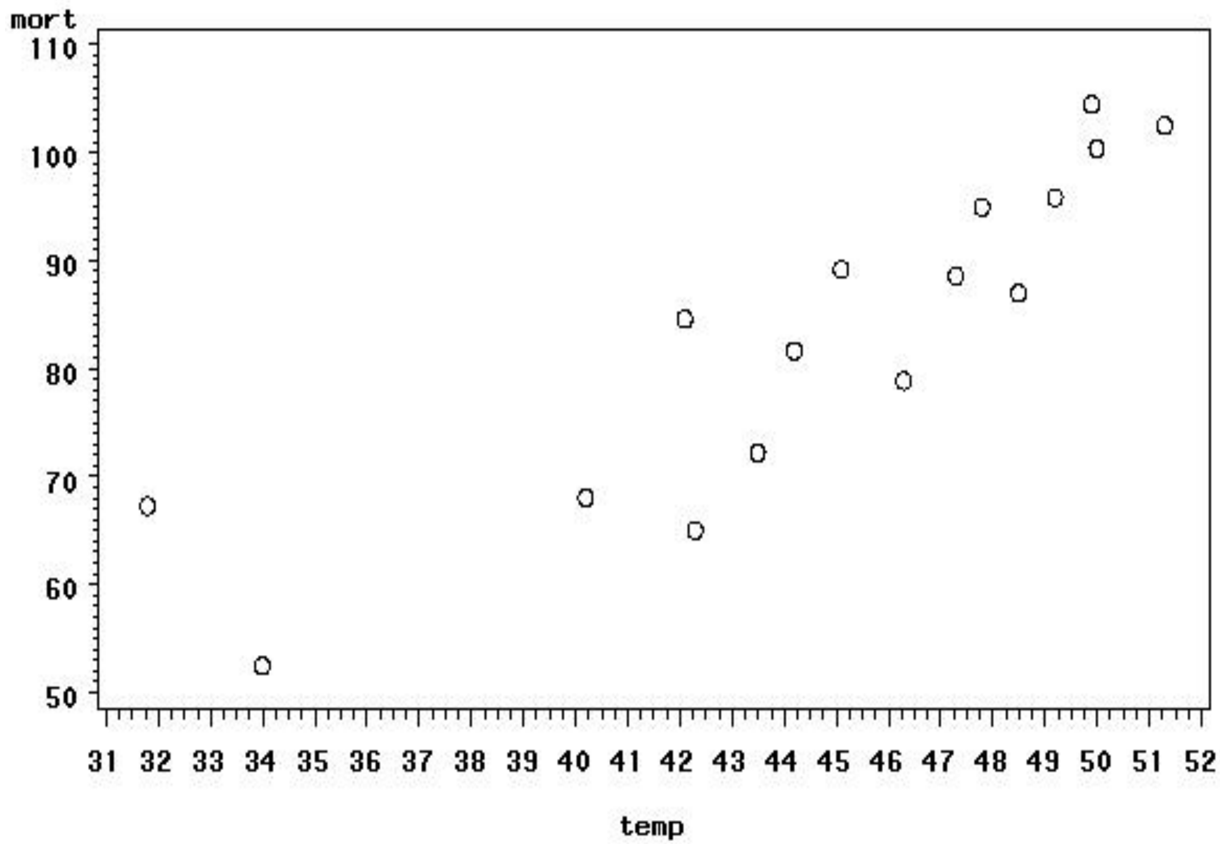
Mort	Temp
102.5	51.3
104.5	49.9
100.4	50.0
95.9	49.2

......

Do it in SAS

```
data breastcancer;
   infile 'breastcancer.dat';
   input mort temp;
symbol1 v=circle;
proc gplot data=breastcancer;
   plot mort*temp;
run;
```

Breast Cancer: Scatter plot



Do it in SAS

```
proc reg
    data=breastcancer;
    model mort=temp;
    run;
```

Mort= -21.8+2.36*temp,

s^2=57.

- What's the interpretation of 2.4? Is it significantly different from zero? How confident are we about this estimate?

Inference for β_1

$$b_1 \sim N(\beta_1, \sigma^2(b_1))$$

$$\text{where}\ \sigma^2(b_1) = \sigma^2 \Big/ \sum (X_i - \bar{X})^2$$

$$t = (b_1 - \beta_1) / s(b_1)$$

$$\text{where}\ s(b_1) = \sqrt{s^2 \Big/ \sum (X_i - \bar{X})^2}$$

$$t \sim \mathrm{t}(n-2)$$

Mathematical Remarks

- Normality of b_1 follows from the fact that it is a linear combination of the Y_i, each of which is an independent normal
- Use results for functions of r.v. to derive its mean and s.d.
- Need independence between b_1 and s^2 to have t-distribution

Notes

- Variance of b_1 smallest among all unbiased estimators of β_1
- Because $\sigma^2(b_1)=\sigma^2/\Sigma(X_i - \overline{X})^2$, we can make this quantity small by making $\Sigma(X_i - \overline{X})^2$ large.

Confidence Interval for β_1

- $b_1 \pm t^* s(b_1)$
- where $t^* = t(1-\alpha/2; n-2)$, the upper $(1-\alpha/2)$ 100 percentile of the t distribution with n-2 degrees of freedom
- $1-\alpha$ is the confidence level

Significance tests for β$_1$

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

$$t = (b_1 - 0)/s(b_1)$$

$$\text{Reject } H_0 \text{ if } |t| \geq t^*, t^* = t(1 - \alpha/2, n-2)$$

$$p - value = \mathrm{P}(|\mathrm{T}| > |t|), \text{ where } \mathrm{T} \sim t(n-2)$$

The book discourages tests in favor of CIs

Breast Cancer Example

Root MSE	7.54466	R-Square	0.7654
Dependent Mean	83.34375	Adj R-Sq	0.7486
Coeff Var	9.05246		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-21.79469	15.67190	-1.39	0.1860
temp	1	2.35769	0.34888	6.76	<.0001

- What's the 99% CI for the regression coefficient of temp ?

Inference for β_0

$$b_0 \sim N(\beta_0, \sigma^2(b_0))$$

$$\text{where}\, \sigma^2(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right]$$

$$t = (b_0 - \beta_0) / s(b_0)$$

$$\text{for}\, s(b_0)\ \text{replace}\, \sigma^2 \text{ by } s^2$$

$$t \sim \mathrm{t}(n-2)$$

Confidence Interval for β_0

- $b_0 \pm t^* s(b_0)$
- where $t^* = t(1-\alpha/2; n-2)$, the upper $(1-\alpha/2)$ 100 percentile of the t distribution with n-2 degrees of freedom
- $1-\alpha$ is the confidence level

Significance tests for β_0

$$H_0 : \beta_0 = 0 \text{ vs } H_a : \beta_0 \neq 0$$

$$t = (b_0 - 0)/s(b_0)$$

$$\text{Reject } H_0 \text{ if } |t| \geq t^*, t^* = t(1 - \alpha/2, n - 2)$$

$$P = \text{Prob } (|\text{z}| > |t|), \text{ where } z \sim t(n-2)$$

Notes

- Usually the CI and significance test for β_0 is not of interest
- If the ξ_i are approximately normal, then the CIs and significance tests are generally reasonable approximations

Do it in SAS

```
/* Ask SAS to give CI */
proc reg data=breastcancer;
   model mort=temp/clb alpha=0.01;
run;
```

- /clb gives CI.
- Confidence level: 1-alpha

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2599.53358	2599.53358	45.67	<.0001
Error	14	796.90580	56.92184		
Corrected Total	15	3396.43938			

Root MSE	7.54466	R-Square	0.7654
Dependent Mean	83.34375	Adj R-Sq	0.7486
Coeff Var	9.05246		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	1	-21.79469	15.67190	-1.39	0.1860	-68.44747	24.85809
temp	1	2.35769	0.34888	6.76	<.0001	1.31913	3.39626

Point Estimation of μ_{Yh}

- $\mu_{Yh} = \beta_0 + \beta_1 X_h$, the mean value of Y for the subpopulation with $X=X_h$
- Point estimate of μ_{Yh}: $\hat{Y}_h = b_0 + b_1 X_h$
- Unbiased: $E(\hat{Y}_h) = \mu_{Yh}$

Distribution of $E(Y_h)$

- $\hat{Y}_h$ is normal (Why?)

- variance $\sigma^2(\hat{Y}_h)$= $$\sigma^2\left[\frac{1}{n}+\frac{\left(X_h-\overline{X}\right)^2}{\sum\left(X_i-\overline{X}\right)^2}\right]$$

Inference for $E(Y_h)$

- Estimate $\sigma^2(\hat{Y}_h)$ by

$$s^2(\hat{Y}_h) = s^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right]$$

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

Inference for $E(Y_h)$

- Confidence Interval: $\hat{Y}_h \pm t^* s(\hat{Y}_h)$
  where $t^* = t(1-\alpha/2, n-2)$

- Significance tests: rarely used in practice

Breast Cancer Example

Root MSE	7.54466	R-Square	0.7654
Dependent Mean	83.34375	Adj R-Sq	0.7486
Coeff Var	9.05246		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-21.79469	15.67190	-1.39	0.1860
temp	1	2.35769	0.34888	6.76	<.0001

- What's the 95% CI of the mean mort of cities with temp=45?

Do it in SAS

- Create new data with X_h
- Use /clm option in proc reg to get CI
- Breast cancer example: give the 95% CI of the mean mortality index for temp=45.

```
/*CI for mean response */
data bc1;
  if _n_ = 1 then temp=45;
    output;
  set breastcancer;

proc reg data = bc1;
  model mort= temp/clm
    alpha=.05;
run;
```

Do it in SAS

```
proc reg data = bc1 noprint;
  model mort= temp/ alpha=.
    05;
  output out=bc2 lclm=lower
    uclm=upper p=yhat;
run;
proc print data = bc2;
  where temp=45;
  var yhat lower upper;
run;
```

- Output CI to bc2
- Print only the CI we want: (80, 88).

Predicting new observations

- Predict $Y_{h(new)} = \beta_0 + \beta_1 X_h + \xi$ for a particular value of $X=X_h$
- Best predictor $\hat{Y}_h = b_0 + b_1 X_h$

  (same as point estimator of $E(Y_h)$)
- Prediction variance:

  $Var(Y_{h(new)})=Var(\hat{Y}_h)+Var(\xi)$

  (larger variance)

Inference for $Y_{h(new)}$

- Estimate prediction variance by:

$$s^2(\text{pred}) = s^2\left[1 + \frac{1}{n} + \frac{\left(X_h - \bar{X}\right)^2}{\sum\left(X_i - \bar{X}\right)^2}\right]$$

- t-distribution:

$t = (Y_{h(new)} - \hat{Y}_h)/s(\text{pred}) \sim t(n-2)$

Breast Cancer Example

Root MSE	7.54466	R-Square	0.7654
Dependent Mean	83.34375	Adj R-Sq	0.7486
Coeff Var	9.05246		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-21.79469	15.67190	-1.39	0.1860
temp	1	2.35769	0.34888	6.76	<.0001

- What's the 95% PI of the mort of a city with temp=45?

Do it in SAS

- Create new data with X_h
- Use /cli option to get PI
- Breast cancer example: 95% prediction interval of a city whose temp=45: (68, 101)

```
proc reg data = bc1;
  model mort= temp/cli alpha=.05;
run;
proc reg data = bc1 noprint;
  model mort= temp/ alpha=.05;
output out=bc3 lcl=lower ucl=upper
    p=yhat;
run;
proc print data = bc3;
  where temp=45;
  var yhat lower upper;
run;
```

Notes

- The standard error (Std Error Mean Predict) given in this output is the standard error of $\hat{Y}_h$, not s^2(pred)
- The prediction interval is wider than the confidence interval

Prediction of mean of m new obs.

- Estimate prediction variance by:

$$s^2(\text{predmean}) = s^2\left[\frac{1}{m} + \frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum\left(X_i - \overline{X}\right)^2}\right]$$

- t-distribution:

  $t = (Y_{h(new)} - \hat{Y}_h)/s(\text{predmean}) \sim t(n-2)$.

- What's the PI for the average mort. Index of two cities whose temp=45?

Confidence band for regression line

- Working-Hotelling CB: $\hat{Y}_h \pm Ws(\hat{Y}_h)$
- where $W^2=2F(1-\alpha; 2, n-2)$
- This gives intervals for *all* X_h
- CI narrower when X_h close to $\overline{X}$

Do it in SAS

```
/*plot confidence band */
symbol1 v=circle i=rlclm99;
proc gplot data=breastcancer;
   plot mort*temp;
run;

symbol1 v=circle i=rlclm95;
proc gplot data=breastcancer;
   plot mort*temp;
run;
```

95% and 99% Confidence band