Lecture 5

- Estimation of subpopulation means
- confidence band for regression line
- prediction intervals
- Analysis of variance table
- General linear hypothesis test
- R^2

Estimation of $E(Y_h)$

- $E(Y_h) = \mu_h = \beta_0 + \beta_1 X_h$, the mean value of Y for the subpopulation with $X=X_h$
- we will estimate $E(Y_h)$ by
- $\hat{\mu}_h = b_0 + b_1 X_h$

Theory for Estimation of $E(Y_h)$

- $\hat{\mu}_h$ is normal with mean μ_h
- (it is an unbiased estimator)
- and variance $\sigma^2(\hat{\mu}_h)$=

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

Theory for Estimation of $E(Y_h)$ (2)

- The normality is a consequence of the fact that $\hat{\mu}_h = b_0 + b_1 X_h$ is a linear combination of Y_i's

Application of the Theory

- we estimate $\sigma^2(\hat{\mu}_h)$ by
- $s^2(\hat{\mu}_h)$= $s^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right]$
- it follows that t= $\frac{\hat{\mu}_h - E(Y_h)}{s(\hat{\mu}_h)}$ ~ t(n-2)
- details for confidence intervals and significance tests are consequences

95% Confidence Interval for $E(Y_h)$

- $\hat{\mu}_h \pm t_c\, s(\hat{\mu}_h)$
- where t_c = t(.975, n-2)
- and $s(\hat{\mu}_h)$ = $\sqrt{s^2(\hat{\mu}_h)}$

```
data a1;
   infile `../data/ch01ta01.dat';
   input size hours;
data a2; size=65; output;
         size=100; output;
data a3; set a1 a2;
proc print data=a3;
proc reg data=a3;
   model hours=size/clm;
run;
```

```
                  Dep Var   Predicted
Obs   size          hours       Value
26      65              .    294.4290
27     100              .    419.3861


Std Error
Mean Predict        95% CL Mean
      9.9176    273.9129    314.9451
     14.2723    389.8615    448.9106
```

Notes

- significance tests can be constructed using this theory
- but they are rarely used in practice

Confidence band for regression line

- $\hat{\mu}_h$ ± Ws($\hat{\mu}_h$)
- where $W^2=2F(1-\alpha; 2, n-2)$
- This gives intervals for *all* X_h
- Boundary values define a hyperbola

Confidence band for regression line

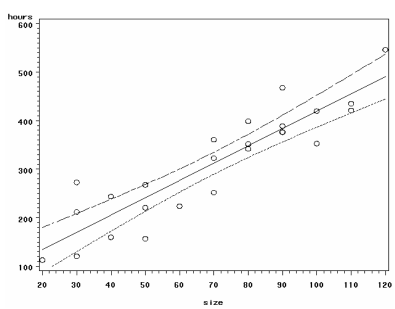
- Theory comes from the joint confidence region for (β_0, β_1) which is an ellipse
- We can find alpha for t_c that gives the same results
- We find W^2 and then find alpha for t_c that will give W = t_c

```
data a1; n=25; alpha=.10;
   dfn=2; dfd=n-2;
   w2=2*finv(1-alpha,dfn,dfd);
   w=sqrt(w2);
   alphat=2*(1-probt(w,dfd));
   tc=tinv(1-alphat/2,dfd);
   output;
proc print data=a1;
run;
```

Obs	n	alpha	dfn	dfd	w2
1	25	0.1	2	23	5.09858

w	alphat	tc
2.25800	0.033740	2.25800

```
data a2;
infile`../data/ch01ta01.dat
';
input size hours;
symbol1 v=circle i=rlclm97;
proc gplot data=a2;
plot hours*size;
run;
```



Prediction of $Y_{h(new)}$

- $Y_h = \beta_0 + \beta_1 X_h + \xi_h$
- Var(Y_h - $\hat{\mu}_h$)=Var Y_h + Var $\hat{\mu}_h$ = σ^2+Var $\hat{\mu}_h$
- S^2(pred)= $$s^2\left[1+\frac{1}{n}+\frac{\left(X_h-\overline{X}\right)^2}{\sum\left(X_i-\overline{X}\right)^2}\right]$$

(Y_h - $\hat{\mu}_h$)/s(pred) ~ t(n-2)

Prediction of Y_h

- **Procedure can be modified for the mean of *m* observations at $X=X_h$**

```
data a1;
   infile `../data/ch01ta01.dat';
   input size hours;
data a2; size=65; output;
         size=100; output;
data a3; set a1 a2;
proc print data=a3;
proc reg data=a3;
   model hours=size/cli;
run;
```

```
                         Dep Var   Predicted
Obs    size               hours       Value
27         100                .    419.3861

Std Error
Mean Predict          95% CL Predict
     14.2723       314.1604    524.6117
```

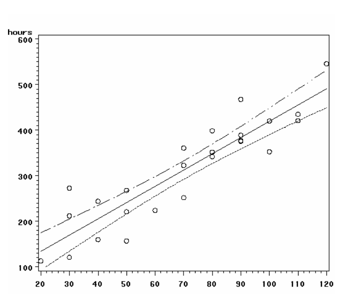
Notes

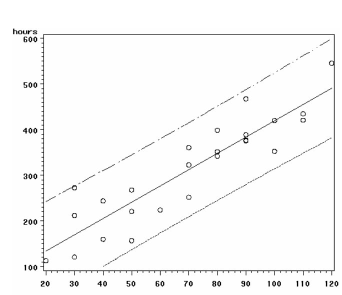
- The standard error (`Std Error Mean Predict`) given in this output is the standard error of $\hat{\mu}_h$, $s^2(\hat{\mu}_h)$, not s^2(pred)
- The prediction interval is wider than the confidence interval

95% Confidence Interval for $E(Y_h)$ and 95% Prediction Interval for Y_h

- $\hat{\mu}_h \pm t_c\, s(\hat{\mu}_h)$
- $\hat{\mu}_h \pm t_c$ s(pred)
- where $t_c = t(.975, n-2)$

```
data a1;
infile
`../data/ch01ta01.dat';
input size hours;
symbol1 v=circle i=rlclm95;
proc gplot data=a1;
plot hours*size; run;
symbol1 v=circle i=rlcli95;
proc gplot data=a1;
plot hours*size; run;quit;
```





Analysis of Variance (ANOVA)

- A way to organize arithmetic
- (Total) variation in Y can be expressed as $\Sigma(Y_i - \bar{Y})^2$
- Partition this variation into two *sources*
  - Model (regression)
  - Error (residual)

ANOVA (Total)

- $SST = \Sigma(Y_i - \bar{Y})^2$
- dfT = n-1
- MST = SST/dfT

ANOVA (Total) (2)

- MST is the usual estimate of the variance of Y if there are no explanatory variables
- SAS uses the term `Corrected Total` for this source
- Uncorrected is ΣY_i^2
- The correction means that we subtract $\bar{Y}$ before squaring: $\Sigma(Y_i - \bar{Y})^2$

ANOVA (Model)

- $SSM = \Sigma(\hat{Y}_i - \bar{Y})^2$
- dfM = 1 (for the slope)
- MSM = SSM/dfM

ANOVA (Error)

- $SSE = \Sigma(Y_i - \hat{Y}_i)^2$
- dfE = n-2
- MSE = SSE/dfE
- MSE is an estimate of the variance of Y taking into account (or conditioning on) the explanatory variable(s))

ANOVA Table

Source	df	SS	MS
Model	1	$\Sigma(\hat{Y}_i - \bar{Y})^2$	SSM/dfM
Error	n-2	$\Sigma(Y_i - \hat{Y}_i)^2$	SSE/dfE
Total	n-1	$\Sigma(Y_i - \bar{Y})^2$	SST/dfT

ANOVA Table (2)

Source	df	SS	MS	F	P
Model	1	SSM	MSM	MSM/MSE	.nn
Error	n-2	SSE	MSE		
Total	n-1				

Expected Mean Squares

- MSM, MSE are random variables
- $E(MSM) = \sigma^2 + \beta_1^2 \Sigma(X_i - \bar{X})^2$
- $E(MSE) = \sigma^2$
- When H_0 is true, $\beta_1 = 0$, E(MSM) = E(MSE)

F test

- F=MSM/MSE ~ F(dfM, dfE) = F(1, n-2)
- When H_0 is false, $\beta_1 \neq 0$ and MSM tends to be larger than MSE
- We reject H_0 when F is large:
- $F \geq F(1-\alpha, dfM, dfE) = F(.95, 1, n-2)$
- In practice we use P values

F test (2)

- When H_0 is false, F has a *noncentral* F distribution
- This can be used to calculate power
- Recall $t = b_1/s(b_1)$ tests H_0
- It can be shown that $t^2 = F$
- So the two approaches give the same P values

```
data a1;
   infile
'h:/STAT512/ch01ta01.txt';
   input size hours;
proc reg data=a1;
   model hours=size;
run;
```

```
                  Sum of    Mean
Source   DF   Squares  Square
Model     1    252378  252378
Error    23     54825    2383
C Total  24    307203

F Value    Pr > F
105.88     <.0001
```

Var	DF	Par Est	St Err	t	Pr>|t|
Int	1	62.36	26.17	2.38	0.0259
size	1	3.57	0.34	10.29	<.0001

General linear test

- A different view of the same problem
- We want to compare two models
  - $Y_i = \beta_0 + \beta_1 X_i + \xi_i$ (*full* model)
  - $Y_i = \beta_0 + \xi_i$ (*reduced* model)
- Compare using SSEs: SSE(F), SSE(R)
- F=((SSE(R) - SSE(F))/(dfE(R) - dfE(F)))/ MSE(F)

Simple Linear Regression

- $SSE(R) = \Sigma(Y_i - b_0)^2 = \Sigma(Y_i - \overline{Y})^2 = SST$
- SSE(F)=SSE
- dfE(R)=n-1, dfE(F)=n-2,
- dfE(R)-dfE(F)=1
- F=(SST-SSE)/MSE=SSM/MSE

R^2 , r^2

- r is the usual (Pearson) correlation
- It is a number between –1 and +1 and measures the strength of the linear relation between two variables
- r^2 = SSM/SST = 1 – SSE/SST
- Explained and unexplained variation

R^2 , r^2

- We use R^2 when the number of explanatory variables is arbitrary (simple and multiple regression)
- R^2 is often multiplied by 100 and thereby expressed as a percent

Source	DF	Sum of Squares	Mean Square
Model	1	252378	252378
Error	23	54825	2383
C Total	24	307203	

F Value	Pr > F
105.88	<.0001

```
R-Square      0.8215 (SAS)
  = SSM/SST
  = 252378/307203

Adj R-Sq      0.8138 (SAS)
  =1-MSE/MST
  =1-2383/(307203/24)
```