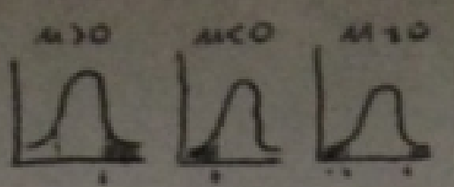
Chapter 11: Test for Means:

1. H_0: $\mu=0$ H_A: $\mu<0$ $\mu>0$ $\mu\neq0$ 2. $t=\frac{\bar{x}}{s/\sqrt{n}}$ $df=n-1$ 3. P-value 4. $p<\alpha$, reject; $p>\alpha$, fail to reject

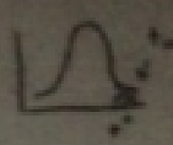
Type I: reject H_0 when its true
Type II: fail to reject when H_A is true

CI: $\bar{x}\pm t^*_{n-1}\left[\frac{s}{\sqrt{n}}\right]$

invT(α,df)
$\frac{\alpha}{2}$



If t_{obs} falls in rejection region, reject H_0



95% confident that true mean lies on interval from (__)
Conditions: Independence (10%, SRS), Nearly normal

Chapter 12: Comparing Two Groups:

Conditions: independent groups
Independent data (10% SRS)
Nearly normal

$\sigma_1 \neq \sigma_2$

Two-sample t-Test:
1. H_0: $\mu_1-\mu_2=0 \rightarrow \mu_1=\mu_2$
H_A: $\mu_1-\mu_2<0$ $\mu_1-\mu_2>0$ $\mu_1-\mu_2\neq0$

$\mu_1 < \mu_2$
$\mu_2 > \mu_1$
$\mu_1 \neq \mu_2$

2-SampT test

2. $t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$

df= smaller of (n_1-1) or (n_2-1)

2-SampTInt

CI: $(\bar{x}_1-\bar{x}_2)\pm t^*_{df}\left[\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$

*95% confident true mean difference will change in this interv
If 0 falls in interval, fail to reject; there is not sig. diff
If 0 does not, reject H_0; there is stat sig difference

Paired t-Test (same population, before/after) *use 1-sample t-test

$\mu_d=\mu_2-\mu_1$
1. H_0: $\mu_d=0$ H_A: $\mu_d<0$ $\mu_d>0$ $\mu_d\neq0$ 2. $t=\frac{\bar{x}_d}{s_d/\sqrt{n}}$ $df=n-1$

3. p-value $\bar{d}=\sum$differences / n 4. P vs. α
$P<\alpha$, reject
$P>\alpha$, fail to reject

CI: $\bar{d}\pm t^*_{n-1}\left[\frac{s_d}{\sqrt{n}}\right]$

Interpreting Hypothesis Tests:
Reject H0 if: outside CI or $p<$ → difference in means is statistically significant
Fail to reject if: inside CI or $p>$ → difference in means is not statistically significant

Pooled t-Test: when $\sigma_1=\sigma_2$ (steps 1, 3, 4 the same) *equal variances*

2. $t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ → SE $df=n_1+n_2-2$ *use 2-samptest→pooled

CI: $(\bar{x}_1-\bar{x}_2)\pm t^*SE$

Rejection Regions: If t_{obs} falls into region, reject H_0 $t^*=$ invT(α,n_1+n

Chapter 14: Inference for Regression:

Conditions: Linearity
Independence
Normal
Equal Spread (residual plot)

Regression Model: $\hat{y}=B_0+B_1x$
b_1=estimate of $B_1=r[S_y/S_x]$
b_0=estimate of $B_0=\bar{y}-b_1\bar{x}$

$\bar{x}=\frac{\sum x}{n}$ $\bar{y}=\frac{\sum y}{n}$

H_0: $B_1=0$
H_A: $B_1\neq0$
Is x/y linearly related?
$t=\frac{b_1}{SE(b_1)}$ $df=n-2$ CI: $b_1\pm t^*_{n-2}SE(b_1)$ SE(b

H_0: $B_0=0$ H_A: $B_0\neq0$
Should line have an intercept?

*95% confident y changes from (__) for each addt'l unit
If 0 falls in interval, fail to reject H_0
If 0 does not fall in interval, reject H_0

Estimations:
What's the average y when $x=x^*=$?
Prediction of all means
$\hat{y}=b_0+b_1x^*$ → estimate μ_y → estimate of pop mean response

CI for $\mu_y=\hat{y}\pm t^*_{n-2}SE(\mu_y)$ $SE(\mu_y)=s_e\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$

CI measures accuracy of mean response
95% confident average y in population is between (__)

Predictions:
What is predicted individual y when $x=x^*$?
prediction of mean of one particular subject
$\hat{y}=b_0+b_1x^*$

$s_e=\sqrt{\frac{\sum(y_i-\hat{y})^2}{n-2}}$ spread a

PI for $\hat{y}=\hat{y}\pm t^*_{n-2}SE(\hat{y})$ $SE(\hat{y})=s_e\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$

PI measures accuracy of individual's predicted value
95% confident value for y in for x^* is between (__)

Chapter 15: Multiple Regression:

$\mu_y=\hat{y}=B_0+B_1x_1+\ldots+B_kx_k+E$ (parameter) $E\sim N(0,\sigma)$

$\hat{y}=b_0+b_1x_1+\ldots+b_kx_k$ residual=$y-\hat{y}$

Tests: 1. H_0: $B_1=0$ H_A: $B_1\neq0$
(tell whether B_1 explains response)

2. H_0: $B_1=B_k=0$ H_A: $B_k\neq0$
(at least 1 predictor is significant)

$t=\frac{b_i}{SE(b_i)}$ $df=n-k-1$ $CI(b_i)=b_i\pm t^*_{n-k-1}$

If we reject H_0, at least 1 variable is stat significant
If we fail to reject H_0, y is not related to all predicted

F-Test ANOVA:

	df	SS	MS	F	p
Regression	k	SSR	MSR	MSR/MSE	p
Error	n-k-1	SSE	MSE	-	-
Total	n-1	SST	MST	-	-

SSR =variation explained by regression eq
SSE =variation not explained by regression eq
SST =variation in response about the mean (SSE+SSR)
Estimate of σ = Se = $\sqrt{MSE}$

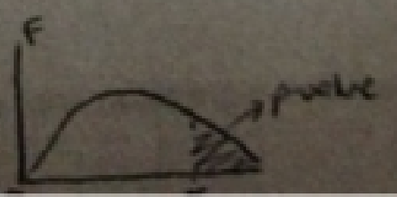
MST= SST / n-1
MSE=SSE / n-k-1
MSR=SSR / k

$R^2=\frac{SSR}{SST}$ if you add insignificant predictors, R^2
$R_{adj}^2=\frac{n-1}{n-k-1}R^2$ if you add insignificant predictors, R_a

R^2 measures % variation of response by predicto

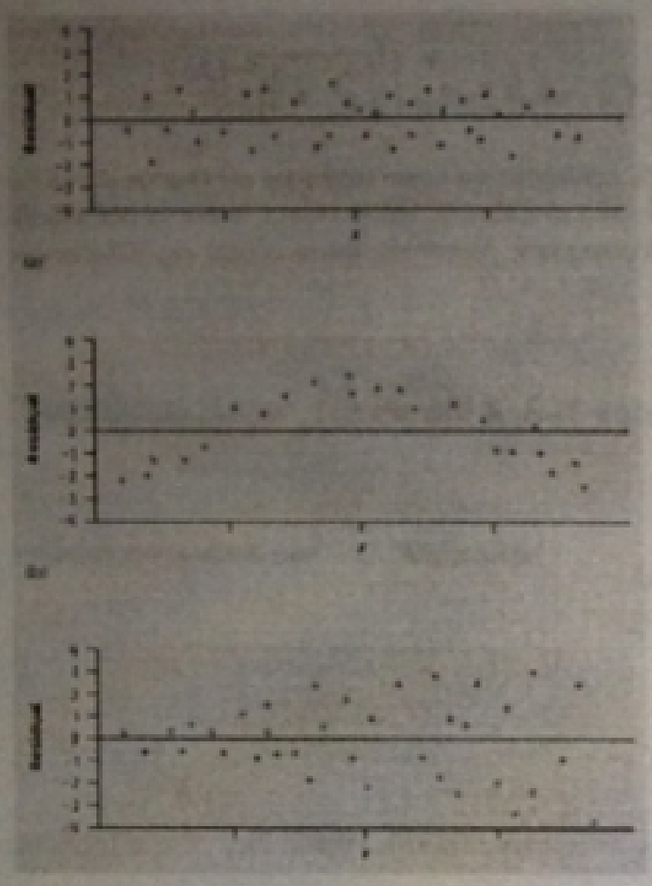
F-Test: overall test of whether ALL predictors in th
model have a slope 0

Compare each p-value to α: if $p<\alpha$ then regression coefficient is significantly different than 0
Each predictor variable has own p-value
σ is estimated by Se, SE about regression line = $\sqrt{MSE}$



hat does regression model say?

- slope: y-value will increase by the slope coefficient for every unit increase
- intercept: negative (meaningless); large p-value=not significance difference
- SE of slope: slopes will vary with SD of SE
- Given s: amount by which y differs from predictors has SD=



Residuals are randomly scattered

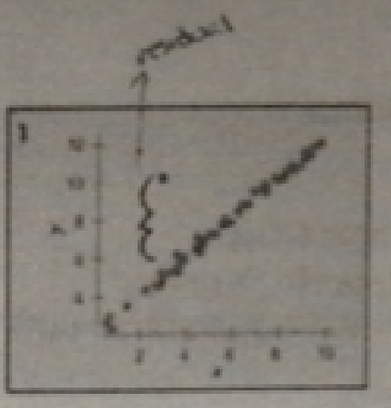
→ good!

Curved pattern

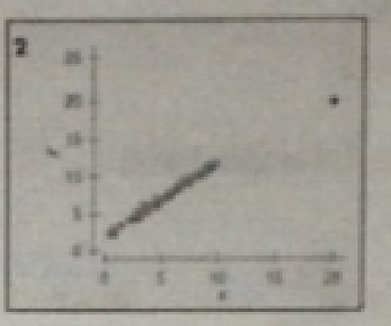
→ the relationship is not linear.

Change in variability across plot

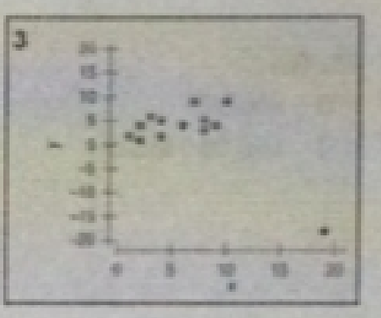
→ σ not equal for all values of x.



- Not high-leverage
- Large residual – far from line
- Not very influential



- High-leverage - far from mean
- Small residual
- Not very influential

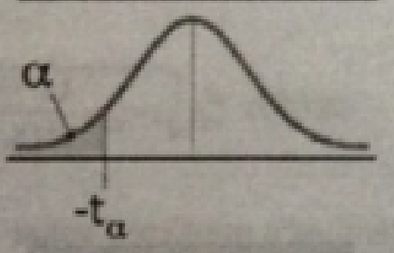
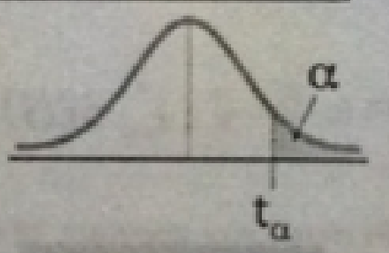
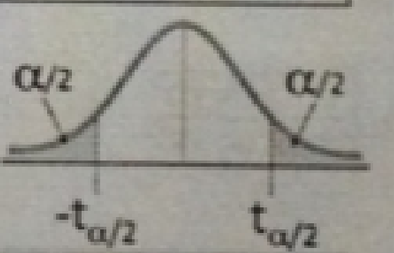


- High-leverage
- Medium residual
- Very influential – changes slope (omitting the red point will change the slope dramatically!)

Hypothesis Testing for Paired Samples

(continued)

Paired Samples (before/after)

Lower tail test:	Upper tail test:	Two-tailed test:
$H_0: \mu_d \geq 0$ $H_a: \mu_d < 0$	$H_0: \mu_d \leq 0$ $H_a: \mu_d > 0$	$H_0: \mu_d = 0$ $H_a: \mu_d \neq 0$
		
Reject H_0 if $t < -t_\alpha$	Reject H_0 if $t > t_\alpha$	Reject H_0 if $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

Where t has n - 1 d.f.

Chapter 6: Scatter Plots & Correlation

Scatter plots:

- **residual**= observed – predicted (distance of a point from the line of regression) negative=overestimate; positive=underestimate
- **correlation coefficient** (r)- strength of a line $r = \sum z_x z_y / n-1$ $z_x = (x-\bar{x}/s_x)$ $z_y = (y-\bar{y}/s_y)$ S_x=stdev of x; S_y=stdev of y
  - does not imply causation
- r^2 – correlation squared; shows % of y explained by x; changes in x explain r% of variations in y
- **lurking variable**- something that affects 2 unrelated variables

Regression:

- $y=b_0+b_1x$
- $b_1=r(S_y/S_x)$
- $Zy=rZx$ → $y-\bar{y}/S_y = r(x-\bar{x}/S_x)$
- **standard error**- $\sqrt{\Sigma}$ sq residuals / n-2 → tells spread about the regression line; sum of residuals=0