Source	DF	Sum of squares	Mean square	F
Model	1	SSM	MSM = SSM / DFM	F = MSM / MSE
Error	n - 2	SSE	MSE = SSE / DFE	
Total	n - 1	SST	MST = SST / DFT	

- This is the ANOVA table for simple linear regression
- Recall our estimate of σ^2 (variance of the residuals) was

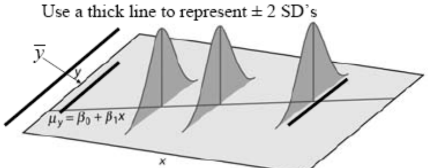
$$\text{MSE} = s^2 = \frac{\text{SSE}}{\text{DFE}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}$$

Visualization of decomposition of variances associated with ANOVA

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

Recall the linkage MST MSM MSE

Use a thick line to represent ± 2 SD's



As β_1 nears 0, MSM becomes small and MSE nears MST

But the key is the size of MSM relative to MSE (noise) 11

Testing the strength of the model

- The main test is whether or not the model works
- In the case of simple linear regression this is the test of H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, which uses the F statistic
  $$F = \frac{\text{MSM}}{\text{MSE}}$$
  has an F distribution with 1 and $n - 2$ degrees of freedom when H_0: $\beta_1 = 0$ is true
- When we get to more complicated models, this will be expanded (e.g. H_0: $\beta_1 = \beta_2 = \beta_3 = 0$)
- If the model variation (MSM) is large compared to error (residual) variation (MSE), then there is strong evidence in favor of the model

"Step-up" approach to regression modeling

- Start with most significant variable in the simple (1 X) regression model (largest r^2, smallest t-test P-value)
  - Wt $r^2 = 0.819$, $F(1,80) = 362$ and $t = -19$ ($P < 0.0001$)
- Vehicle weight (Wt) enters the model in the 1st step up
- To the model already containing Wt, add each of the remaining X variables one at a time looking at the additional contribution (r^2 and P-value)
  - Cab $r^2 = 0.820$, F = 180 and $t = -0.46$ ($P = 0.65$)
  - HP $r^2 = 0.824$, F = 184 and $t = -1.40$ ($P = 0.166$)
  - Speed $r^2 = 0.829$, F = 192 and $t = -2.18$ ($P = 0.033$)
- Addition of Speed to the model gives the largest r^2 and the most significant t-test result
- Speed is 2nd predictor variable to enter the step-up model

11

"Step-up" approach to regression modeling

- In the 2nd step the model contains Wt and Speed
- To the model already containing Wt and Speed add each of the remaining X variables one at a time looking at the additional contribution (r^2 and P-value)
  - Cab $r^2 = 0.835$, F = 132 and $t = -1.64$ ($P = 0.106$)
  - HP $r^2 = 0.873$, F = 178 and $t = 5.13$ ($P < 0.0001$)
- Addition of HP to the model gives the largest r^2 and the most significant t test result
- Does HP provide significant ($P < 0.05$) additional prognostic information?
- Yes ($P < 0.0001$), so we continue the step-up process
- HP is 3rd predictor variable to enter the step-up model

12

"Step-up" approach to regression modeling

- In the 3rd step the model contains Wt, Speed & HP
- To the model already containing Wt, Speed & HP add the remaining X variable to see if it is needed in the model
  - Cab $r^2 = 0.873$, F = 133 and $t = -0.69$ ($P = 0.50$)
- Cab does not provide significant additional prognostic information so the final model contains only MPG (y) plus Wt, Speed, HP [plus the constant]
- The final model is
  $\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$ or
  $\hat{y}(MPG) = 194.1 - 1.92(Wt) - 1.32(Speed) + 0.41(HP)$
- Check the regression assumptions for the final model

13

One-way ANOVA table

Source	DF	Sum of squares	Mean square	F
Groups	I - 1	SSG	s_B^2 = SSG / DFG	F = MSG / MSE
Error	N - I	SSE	s_W^2 = SSE / DFE	
Total	N - 1	SST	SST / DFT	

- The F statistic tests if there is a difference among the I population means
- MSE is still our estimate of σ^2 (variance of the residuals)

Two-way ANOVA table

Source	DF	Sum of squares	Mean square	F
A	I - 1	SSA	SSA/DFA	MSA/MSE
B	J - 1	SSB	SSB/DFB	MSB/MSE
AB	(I-1)(J-1)	SSAB	SSAB/DFAB	MSAB/MSE
Error	N - IJ	SSE	SSE/DFE	
Total	N - 1	SST	SST/DFT	

- F tests for main effect A, main effect B and interaction AB (note all are divided by MSE)
- MSE is still our estimate of σ^2 (variance of the residuals)

Two-way ANOVA – the model

- SRSs of size n_{ij} from each of I x J normal populations
- The population means μ_{ij} may differ, but all populations have the same SD - σ
- Let x_{ijk} be the k^{th} observation from the population having factor A at level i and factor B at level j
- The two-way ANOVA model is
  $$x_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$
  for $i = 1, 2, .., I$ and $j = 1, 2, .., J$ and $k = 1, 2, .., n_{ij}$ where the deviations ε_{ijk} are assumed to be ~ N(0,σ)
- Model parameters are μ_{ij} and σ

Two-way ANOVA

- In one-way ANOVA the sum of squares are decomposed
  SST = SSG + SSE
- In two-way ANOVA the sum of squares are decomposed
  SST = SSA + SSB + SSAB + SSE where
  - SSA - main effect for A
  - SSB - main effect for B
  - SSAB - AB interaction
  - SSE - usual term for error (residuals)
- Degrees of freedom are now partitioned
  DFT = DFA + DFB + DFAB + DFE
  (N - 1) = (I - 1) + (J - 1) + (I - 1)(J - 1) + (N - IJ)
- Mean squares (MS) are formed the usual way

Let μ_d be the mean of the difference in admissions between Friday 13th and Friday 20th
Note: this is a 1-sided test of hypothesis
H_0: $\mu_d = 0$ versus H_a: $\mu_d > 0$
$t = (\bar{x}_d - \mu_o) / [s_d/\sqrt{n}] = (3.4 - 0) / [4.3/\sqrt{10}] = 2.50$
The degrees of freedom are n - 1 = 10 – 1 = 9
Reject H_0 if $t > t_{0.05,9} = 1.833$
$P(t > 2.50)$ is between 0.01 and 0.02 so $0.01 < P < 0.02$
We reject H_0
We conclude that hospitalizations due to accidents were higher on Friday the 13th as compared to Friday the 20th.

Let X be the number of times there are more hospital admissions due to accidents on Friday the 13th as compared to Friday 20th
H_0: $p = 0.5$ versus H_a: $p > 0.5$
Note: this is a 1-sided test of hypothesis
We can use the binomial distribution [B(10,0.5)] to conduct this sign test.
We observed X = 8.
The P value would be the sum of the probabilities P(X = 8, 9 or 10).
Because this is a binomial distribution, we can calculate the P value using the binomial distribution formula.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ or in this case } P(X = k) = \binom{10}{k}(1/2)^{10}$$

$$P(X = 10) = \binom{10}{10}(1/2)^{10} = 1(0.0009766) = 0.0009766$$

$$P(X = 9) = \binom{10}{9}(1/2)^{10} = 10(0.0009766) = 0.009766$$

$$P(X = 8) = \binom{10}{8}(1/2)^{10} = 45(0.0009766) = 0.043947$$

Thus P(X = 8, 9 or 10) = 0.043947 + 0.009766 + 0.0009766 = 0.0547
So P = 0.0547
We have insufficient evidence to reject H_0
We have insufficient evidence to conclude that hospitalizations due to accidents were higher on Friday the 13th as compared to Friday the 20th.

The t-test conducted in part (a) involved a small sample size and thus depended upon a normal distribution of the data. The normal probability plot showed that the data were not normally distributed. A large outlier value (13) appears to have dominated the results of the t-test. The sign test is a non-parametric procedure and doesn't need any such assumptions. Therefore the sign test result should be trusted more in this instance.