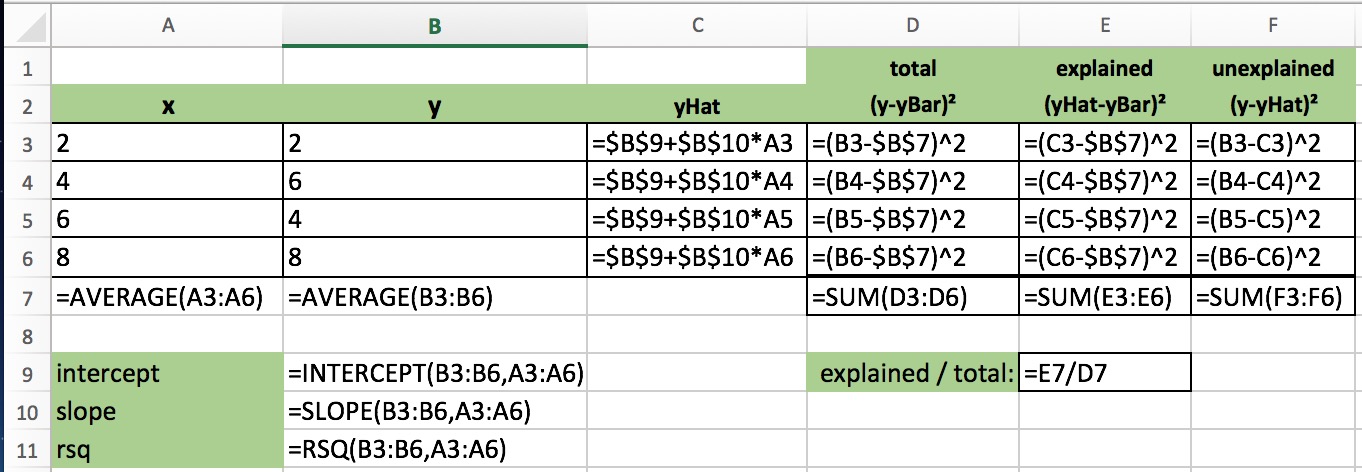
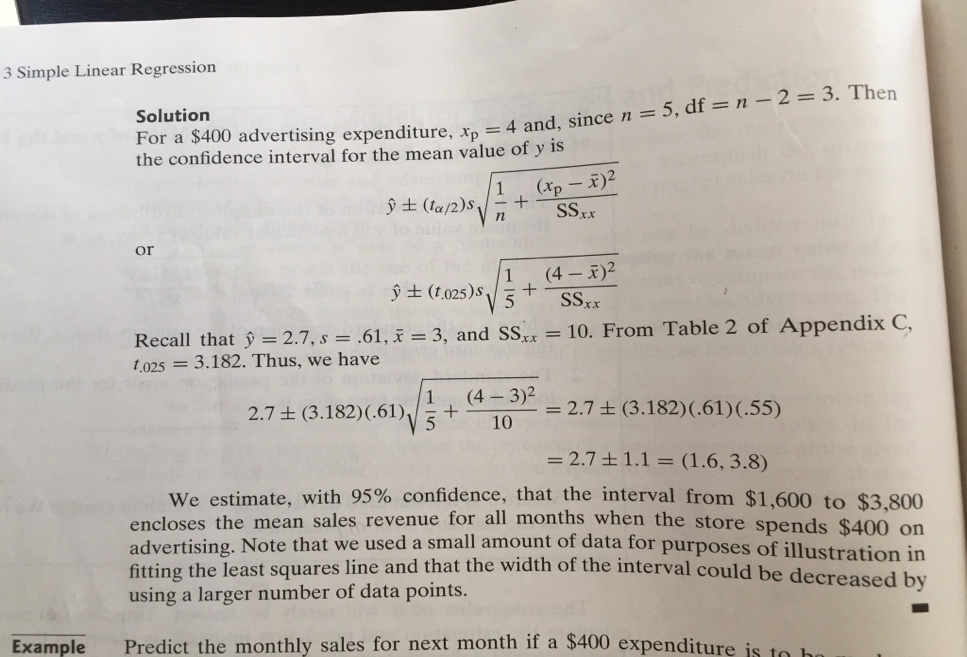
**Be able to perform all of the calculations contained in slides 18 and 27 of week 2 notes.**

**Given partial SAS output, find the confidence interval or complete a test of hypothesis for difference in means using pooled variance.**



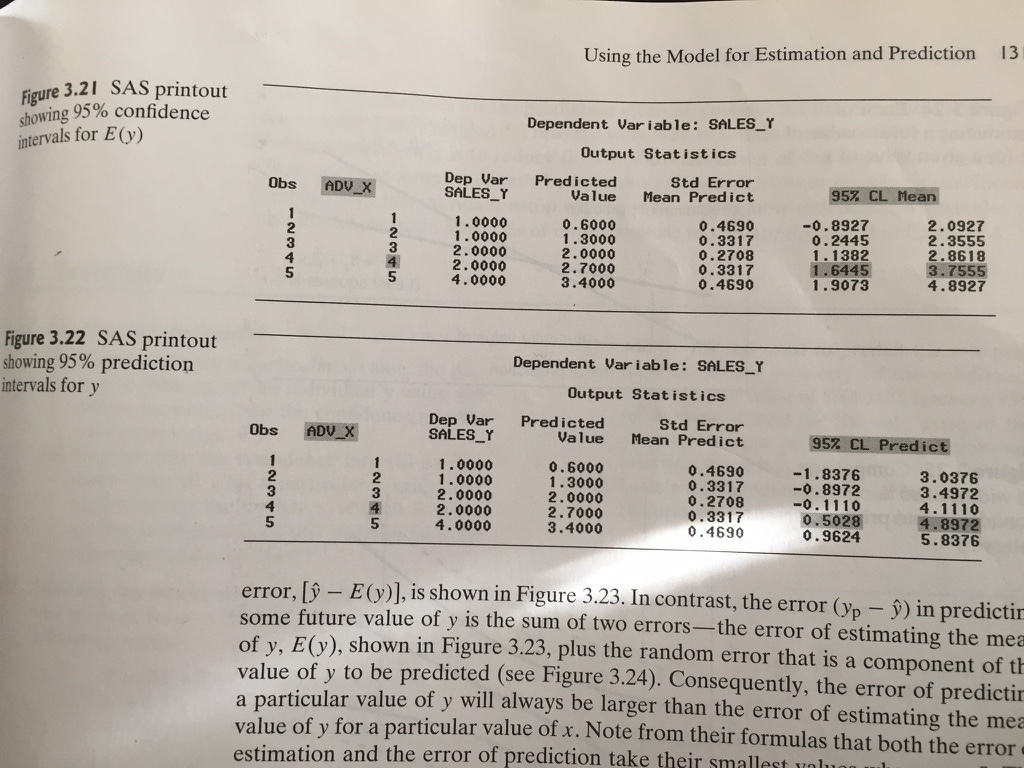
1. H0: = 0 versus HA: ≠ 0

2. where

3. alpha = 0.05

4. Critical value(s) / Rejection region (df = n – (k+1)) look-up table

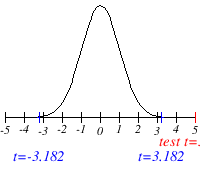
5. Draw the sample

****6. Calculate the observed value of the test statistic

7. State the conclusion

8. Find the p-value.

**Given partial R output, conduct test of hypothesis on beta1. Must identify the critical value, rejection region, and rejection rule, observed value of test statistic, and conclusion.**



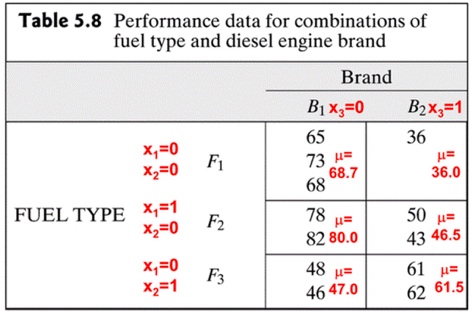
5.

* 1. H0: β1 = 0 versus HA: β1 ≠ 0
  2. t =
  3. alpha = 0.05
  4. Critical value(s) / Rejection region (df = n -2)
     + Rule: Reject H0 if t < -cv or t > +cv

6. Calculate t and shade in on number 5 graph

7. Conclusion: either *Reject H0* [meaning a linear relationship between x & y] or *Fail to reject H0*

**Given one quantitative variable and two qualitative variables and SAS coefficients, complete a table of means by category. Determine if there is an interaction effect and, if so, identify the matching interaction plot (some labels and values hidden).**

If the p-value is lower than the α with an F-statistic so that concludes that the x-variables contribute to the prediction the response (y).

H0: β4 = β5 = 0 versus HA: At least one of β4 and β5 differ from 0

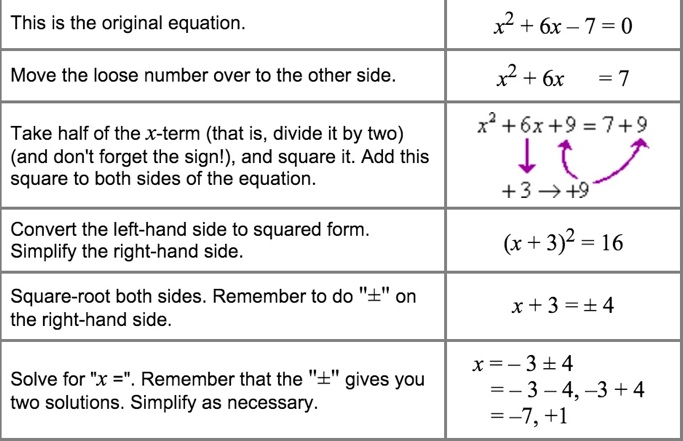
Test Statistic: = df for error for complete model

**Where** Sum of Squares Error for reduced (main effects model)

= Sum of Squares Error for complete (with interaction model) [β4x1x3 & β5x2x3]

g = # of β’s tested => Numerator df for the F statistic

Interaction plots that rise and fall together have interaction

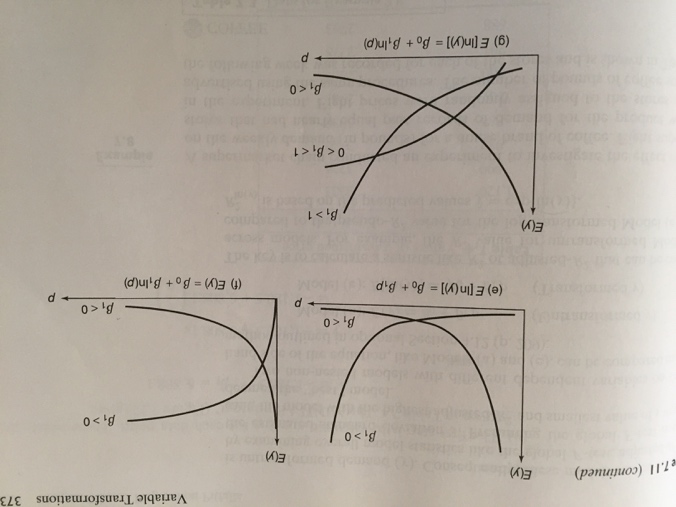
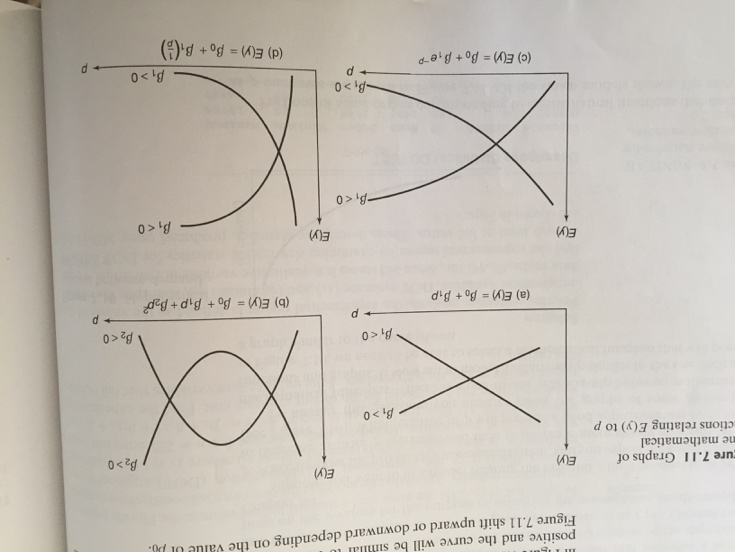
**Given a quadratic model, and output from R, rewrite the model using the complete-the-square method. Find the vertex (turning point) of the parabola (a graph is given but one axis is unlabeled).**

(x + 3)2 = 16 It gives us the **vertex** (turning point) of x2 + 4x + 1: **(-3, -16). And the R code will have the (Intercept)**

**Given R code and output for forward, backward, stepwise, and best n subset (using adjr2) selection, determine the best model for each, and explain why. Where appropriate, identify which variables entered or left the model and when.**

**Forward** – starts with best predictor and adds predictors if they improve the model (LOW AIC is best). **Backward** – elimination and starts with full model and keeps on dropping off the least significant. Once dropped can’t be added back in. **Stepwise** – starts empty, and goes forward or backwards. **Best n subset** – Chooses the model with the highest adjr2 as it explains the highest percentage of variation in y. only increases if the added x variable has significant association with the response variable. increase for any independent variable added to the model even when x-variable has very little predictive power.

**Given a plot, and several attempts at transforming the y to create a linear model, which model transformation is best? Given an x, find the predicted value (transformed and untransformed) for y. Example uses R.**



**Residual analysis: what are homoscedasticity, heteroscedasticity, outliers, influential values, and autocorrelation of residuals? What are their causes? How are they detected?**

**Homescedastic** – constant variance for all levels of independent variables (an assumption necessary for the validity of regression inferences)

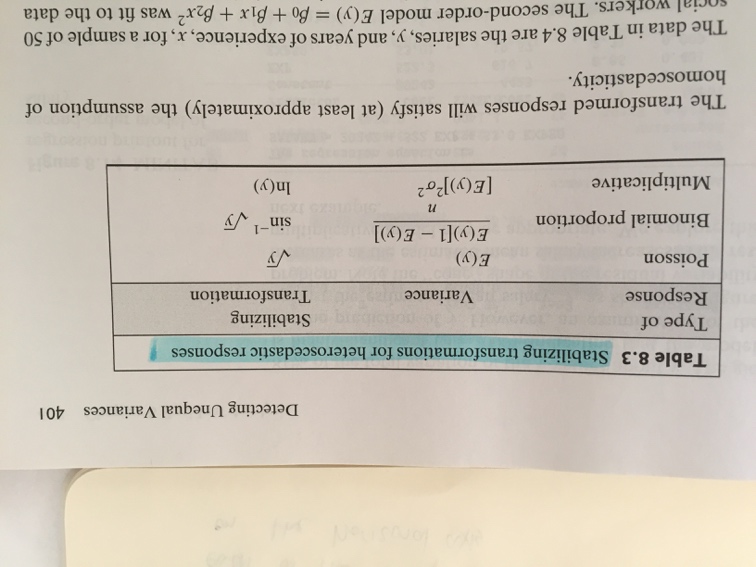
**Heteroscedastic** – unequal variances [**causes**: variance of response y is a function of its mean E(y) like a count data (# of sick days) == Poisson distribution; when response y is a proportion (percentages) generated by == binomial experiments]

* + - **Solutions** – Stabilize the variances by transforming the y (response) by applying either square root or logarithmic; Residual plot will **detect** curvature if it exists but not reveal the appropriate transformation

**Influential Observations** – if removed causes significant changes to the betas / **Outliers** – An observation with a residual that is larger than 3s (stdev);

* + - **Solutions** – we should attempt to identify the cause of outliers: they are often the result of errors (aka assignable errors) and can be excluded from the analysis; Techniques: Leverage, Jackknife, Cook’s Distance.

**Auto-Correlation of residuals** – correlated with itself – usually occurs in time-series data where t time :: t+1 and the model will look better than it actually is; **Solution** – Durbin-Watson test

****