PBHS 32400/Stat 22400
Applied Regression Analysis
SAMPLE questions - Note: actual exam may be longer

1. In each of the following, circle ALL statements that are true based on the information given

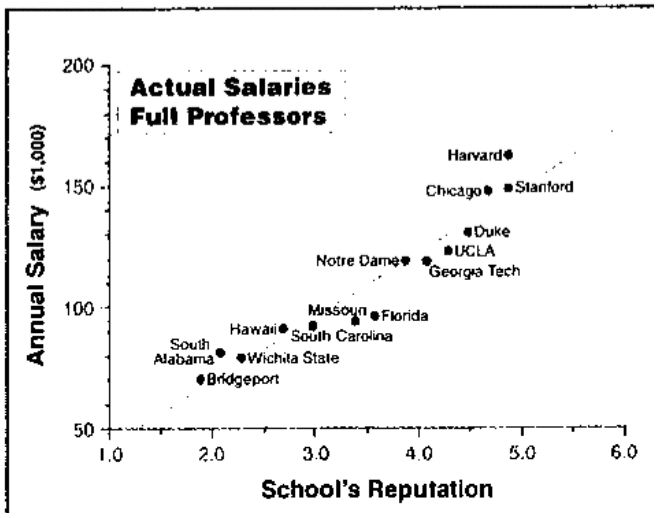
   (a) In a simple linear regression (SLR) of Y being predicted by predictor X, if one obtains an R^2 of 0.25, this implies that
      i. the correlation of X and Y is 0.5 or -0.5
      ii. The slope of the resulting line will be positive
      iii. The F-test for overall model fit will be statistically significant

   (b) From the SLR analysis, I obtain a 95% confidence interval for β_1.
      i. My point estimate $\hat{\beta}_1$ is inside this confidence interval, therefore I do not reject the null hypothesis of $\beta_1 = 0$
      ii. The confidence interval provides all the estimates of the population slope β_1 that are consistent with the data at 95% confidence
      iii. The confidence interval is constructed using a t_{n-2} statistic, because the sampling distribution of $\hat{\beta}_1$ follows a t distribution and $n-2$ is the correct degrees of freedom for SLR with n observations

   (c) Based on simple linear regressions, the R^2 of Y predicted by X_1 is 0.2, the R^2 of Y being predicted by X_2 is 0.4. Then the R^2 for the multiple linear regression with Y being predicted by X_1 and X_2
      i. Will be between 0.4 and 0.6.
      ii. Is irrelevant, since one would simply use X_2 as the best predictor
      iii. Cannot be determined at all from the information given

   (d) If an observation has a residual = 0,
      i. it indicates that the fitted value $\hat{y}_i$ is equal to the observed y_i.
      ii. The leverage for this observation is 0.
      iii. It is suggested that since some values can be predicted perfectly, the R^2 for the model must be very high

2. (8pt) The following questions are based on an analysis of the relationship between university academic reputation and salary of professors .



(a) The regression line has (approximately) estimated coefficients $\hat{\beta}_0 = 15$ and $\hat{\beta}_1 = 25$. What can we say about the increment in salary per 1/2 point in school reputation?

(b) The correlation coefficient associated with the above plot is $r = 0.96$. Provide the R^2 for the model and explain what it means.

(c) Based on the above model (in a.), what is the predicted mean salary for Duke (to the best of your approximation)?

(d) Based on the above model (in a.), what is the predicted mean salary difference between Notre Dame and South Carolina (to the best of your approximation)?

Answer: This would be 25*(difference in reputation scores)

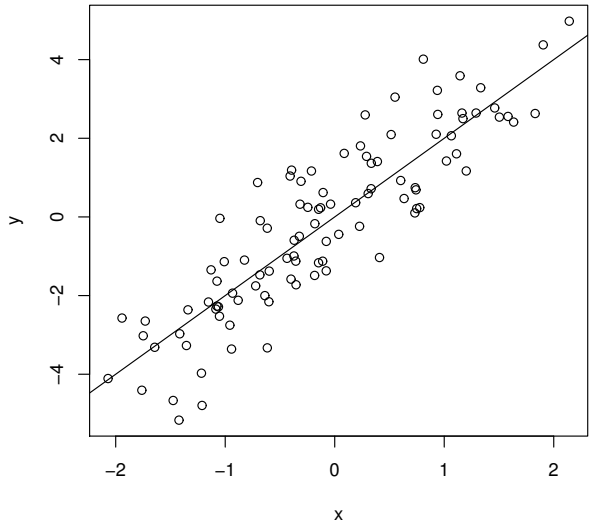
or, compute salary for each and take the difference (the intercepts are the same, so ithis s equivalent to doing as above)

3. (7pt) Residuals: In least squares regression model theory, we denote the random deviations from the predicted line residuals, *errors*, or ϵ_i

(a) What key properties (name two) do we assume about the ϵ_i in the regression model?

(b) After fitting the model, what is the quantity that represents the ϵ_i value for each observation? Either a) draw a picture of how these quantities relate to the fitted line or b) name a property that these quantities will always have.

4. (9pt) Below is a scatter plot of the response variable Y versus the predictor X. The straight line is our fitted SLR line.



(a) $\hat{\beta}_1 =$_____
(note: this can be approximated from basic algebra, need not do least squares)

(b) Which of the followings will be **changed** if we change the units of Y (for example, convert values from kilograms to pounds).
(Circle ALL RESPONSES that apply.)

i. The variance of Y.
ii. The adjusted R-squared.
iii. The estimated coefficients (βs).
iv. The confidence interval for the coefficients.
v. The t-statistics for testing whether a coefficient equals zero.
vi. The degrees of freedom of the overall F-test.
vii. The sum of squared residuals.

Answer

For this, we need to think about each quantity. Quantities related to inference generally do not change, because these are statistical tests formed as ratios.

i changes because Y has a different mean, different values around it

ii R^2 does not change

iii β changes with Y as it is slope on new Y scale

iv CI changes, has to encompass new β

v t-statistic does not change - ratio of new beta to new variance, both are in same units

vi df does not change, the model is the same wrt predictors, which determines df

vii This is sum of residuals on new Y scale, changes

This is a bit hard to think through at first, you can look at a data example to confirm. Just pick a simple model from our datasets, and divide y by, say, 10. Compare the output.

5. **Analysis of Diamond Price Data** The models below concern modeling diamond prices according to elements of the '4 C's' - carat weight, color, clarity, and cut. The response variable is cost in U.S. dollars. Predictors here are carat weight, a numeric color variable (that could be considered either as ordinal or as categories, and cut quality in categories (although these are logically ordered too). Address the following questions about the analysis.

   (a) Carat weight and color (coded as ordinal numeric starting at 1 for near colorless) were evaluated, yielding the following model.

```
. reg price carat ncolor

      Source |       SS           df       MS      Number of obs   =        55
-------------+----------------------------------   F(2, 52)        =    165.56
       Model |   287987020         2   143993510   Prob > F        =    0.0000
    Residual |  45226298.4        52  869736.508   R-squared       =    0.8643
-------------+----------------------------------   Adj R-squared   =    0.8591
       Total |   333213319        54  6170617.01   Root MSE        =     932.6

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       carat |   7607.153   432.4639    17.59   0.000     6739.351    8474.956
      ncolor |  -453.0912   76.85718    -5.90   0.000    -607.3164    -298.866
       _cons |  -486.6812   321.9613    -1.51   0.137    -1132.744    159.3813
------------------------------------------------------------------------------
```

   i. write out the estimated model

   ii. What is the increase in price per carat?

   iii. What happens to price as color goes up the scale?

   iv. What is the price for a .75 carat diamond of color= 1? What is the price for color = 7?

v. Show how the confidence interval on the coefficient for carat is computed

Confidence intervals on β are like CI on a mean. The set-up is

$$(\hat{\beta} - t_{n-p-1,\alpha/2} * se, \hat{\beta} + t_{n-p-1,\alpha/2} * se)$$

Where se is the estimated standard error of the β, and the t-stat is the t distribution reference value with n-p-1 df and the desired significance level (95% is most popular). See notes or book.

(b) We then add a carat by color interaction effect, as follows.

```
. gen carat_by_color = carat*ncolor

. reg price carat ncolor carat_by_color

      Source |       SS           df       MS      Number of obs   =        55
-------------+----------------------------------   F(3, 51)        =    130.24
       Model |   294742323         3  98247440.9   Prob > F        =    0.0000
    Residual |  38470995.8        51  754333.251   R-squared       =    0.8845
-------------+----------------------------------   Adj R-squared   =    0.8778
       Total |   333213319        54  6170617.01   Root MSE        =    868.52

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       carat |   9842.352    848.588    11.60   0.000      8138.74    11545.96
      ncolor |   62.09239   186.4425     0.33   0.740    -312.2065    436.3913
carat_by_color |  -615.2261   205.5862    -2.99   0.004    -1027.958   -202.4946
       _cons |  -2126.921   624.7621    -3.40   0.001    -3381.183   -872.6586
------------------------------------------------------------------------------
```

i. write out the estimated model

Answer

$\widehat{price} = -2126 + 9842 * carat + 62 * color - 615 * carat * color$

ii. Explain the meaning of the interaction term, i.e., what does it do in the model?

Answer: The interaction term allows the effect of carat weight to be different at each level of color. The model in (a) without the interaction terms requires that the per carat effect is the same within each color level.

iii. What is the slope per carat for color= 3?

Answer This one is a bit harder than what we did in class w/just categorical by continuous interaction, but comes out the same

Another way to write above model would be:
$\widehat{price} = -2126 + 62 * color + carat * (9842 - 615 * color)$

Here, it can be seen that the effect (i.e., slope) of carat weight takes on a specific value depending on which color is plugged in.

Another way to see would be to fix color at 3, and calculate the actual price for two diamonds, one with 1 carat and one with 2 carats. The difference between them would come out to (9842 - 615*3). This is the change due to a one carat increment, or in other words, the slope.

(c) A categorical index set is needed for diamond cut, with 4 rating categories (fair, good, very good, excellent) . Create appropriate indicator variables to represent cut as a nominal variable in the model

Answer: We have 4 cut levels, so we need three indicators as follows:

cutgood = 1 if cut=good, zero otherwise

cutverygood = 1 if cut= very good, zero otherwise

cutexcellent = 1 if cut= excellent, zero ptherwise

Each of these will be an 'offset' from the baseline or reference category, which is fair cut. One could run a regression with just these indicators, and this would be equivalent to an ANOVA comparing mean price by cut quality.