Name: ________________________

ID: ________________________

Quiz section or time: ________________________

12 + 25

Stat/Math 390, Winter, Test 2, Feb. 13, 2015; Marzban

Same deal as before ...

Points

Lect 9, p.4

1 **1.** Suppose two continuous variables are sinusoidal in time. Then, their scatterplot will generally appear as

a) sinusoidal   b) cosinusoidal   c) cigar-shaped cloud   d) elliptical/spiral

1 **2.** Suppose data suggest a linear relationship between two continuous variables x, y, with correlation coefficient r. When predicting y from x, the prediction errors

a) will depend on r, only if the relationship is not causal.

b) will depend on r, regardless of causality.

c) will not depend on r, if the relationship is causal.

d) will not depend on r, regardless of causality.

hw 2

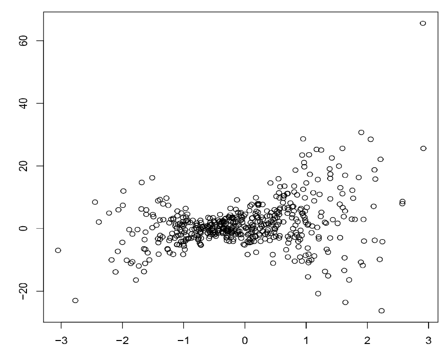
1 **3.** Consider the following scatterplot. Which is correct?

a) If the axes are response versus predictor, then an interaction term is warranted.

b) If the axes are one predictor vs. another predictor, then an interaction term is warranted.

c) If the axes are response versus predictor, then there exists collinearity.

d) If the axes are one predictor vs. another predictor, then there exists collinearity.



Lect 11, p.2

1 **4.** Given data on 10 people on two continuous variables x, y, which of the following is/are true?

a) A correlation coefficient can be computed only if x and y have the same physical units.

b) Regression can be performed only if x and y have the same physical units.

c) A regression equation developed on the data can be used to predict the y value of a "new" person (not among the 10).

d) A correlation coefficient computed from the data can be used to compute the correlation coefficient for a new person (not among the 10).

9 Z5, Lect 13

1 **5.** Circle the true statement. In a polynomial regression involving one predictor x and one response y, with correlation coefficient r between them, generally

a) $R^2 \geq r^2$ b) $R^2 \leq r^2$ c) $R^2 = r^2$ d) None of the above

3.17, Sample Test

1 **6.** In a problem involving data on two continuous variables x, y, grouping the data and averaging the x and y values within each group, will generally lead to a (circle all correct answers)

a) lower r b) comparable r c) higher r d) smaller β e) comparable β f) larger β

0.5 penalty   OK (regression effect)

Lect 15, p.3

1 **7.** A multivariate regression model of data has led to the following "best" model: $y = 2.0 + 3.1x_1 + 2.4x_2 + 1.0x_1x_2$. Which conclusion(s) is/are warranted?

a) There exists collinearity.

b) A change of 1 unit in x_1 leads to an average increase of 3.1 units in y, if there is no collinearity between x_1 and x_2.

c) A change of 1 unit in x_2 leads to an average increase of 3.4 units in y, if there is no collinearity between x_1 and x_2.

d) None of the above.

2

8. Recall that overfitting occurs when the model has many more parameters than the number of independent cases in the data. Therefore, this situation can occur if (circle all correct answers)
a) the model is polynomial regression with very high order
b) the model is multiple regression with many predictors
c) the model is multiple regression with many interactions
d) there exists collinearity

Lect 15 p.6

1 **9.** Suppose there are three different voltage settings on an electric device designed to supply power power to 4 different types of LEDs, each of which comes in two different colors. The best tool for understanding the relationship between voltage, LED type, and color is
a) 3d scatterplot b) 3d contingency table c) multiple regression d) comparative boxplots

Lect 16

1 **10.** In spite of its confusing name, a "sampling distribution" is a
a) distribution of a sample statistic
b) distribution of a distribution parameter
c) histogram of a sample statistic
d) histogram of a distribution parameter

Lect 16

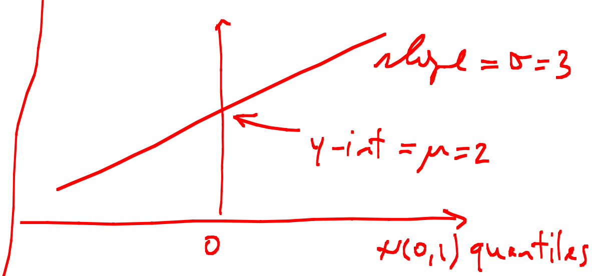
1 **11.** The formulas we have derived ($\mu_{\bar{x}} = \mu_x \; \sigma_{\bar{x}} = \sigma_x/\sqrt{n}$) are correct
a) if the population is normal
b) even if the population is non-normal, but n is large
c) if the population distribution is symmetric
d) None of the above

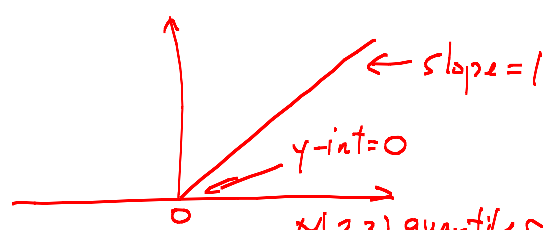
Lect 9 Lab

12. Draw the qqplot corresponding to the following situations; if the qqplot is a straight line, SPECIFY the (y-intercept, slope); if a qqplot is NOT possible, EXPLAIN why not.
Data are from N($\mu = 2, \sigma = 3$), and the x-axis of the qqplot shows quantiles of

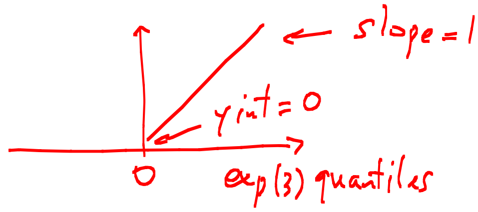
$\sim 1 + 2$ **a)** the standard normal: **b)** N($\mu = 2, \sigma = 3$):

$\frac{9}{2}$





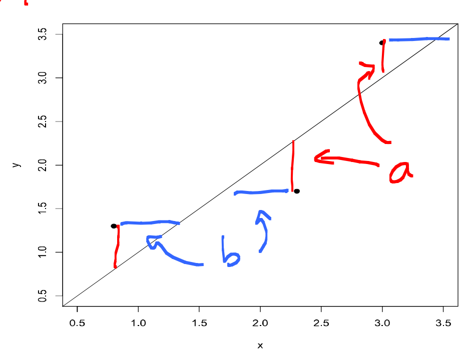
~ 2 **c)** Data are from exp($\lambda = 3$), and the x-axis of the qqplot shows quantiles of exp($\lambda = 3$):



Sample test/class discussion

~ 1 **13.** Consider the data (3 circles) shown.
a) Suppose the line is the OLS fit $y = \alpha + \beta x$. Draw (or show) the errors whose sum of squares is minimized.
~ 2 **b)** On this same figure, draw (or show) the errors whose sum of squares will be minimzed if the line is the OLS fit $x = \alpha' + \beta' y$.
Clearly distinguish your answers to parts a and b.
Do NOT explain.

2

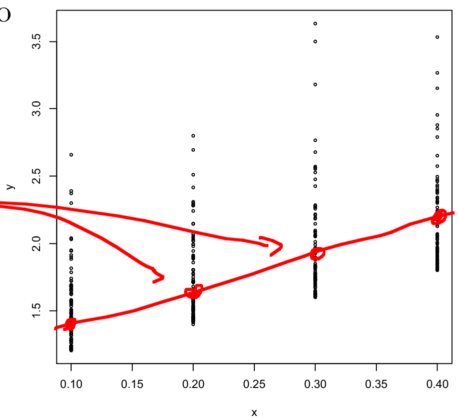


~2 9z4 2*5

14. Without doing any calculations, draw the OLS **LINE** fit to this data. Explain the reason for the line you draw.

OLS fits are designed to go thru the average y-values, at each x.

Given the higher density of data at lower y-values (at each x), the avg is on the lower end of y-values.



15. In a simple linear regression fit to three cases, suppose the residuals are 1.0, -3.0, 2.0, and the standard deviation of the y values is 3.

~2 **a)** What is the typical error committed by this fit? Show work.

typical error

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = (1)^2 + (-3)^2 + (2)^2 = 14 \Rightarrow s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{14}{3-2}} = \boxed{\sqrt{14}}$$

4

~2 **b)** Compute R^2, and interpret it in "English."

$$SS_{expl} = \underset{(n-1)s_y^2}{\underset{\shortparallel}{SS_T}} - SS_E = (3-1)(3)^2 - 14 = 18 - 14 = 4$$

$$R^2 = SS_{expl} / SS_T = \frac{4}{18} = \boxed{\frac{2}{9}} \sim .22$$

[About 22% of the variability in y is due to the variability in x.

hw-P,V

~3 **16.** Find the normal equations for the OLS estimates of α and β for the model $y = \alpha + \beta x^3 + \epsilon_i$. It's important that you do it this way: Start from the expression for SSE, and differentiate, etc.

$$SSE = \sum_i \epsilon_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \alpha - \beta x_i^3)^2$$

$$\frac{\partial}{\partial \alpha} : \sum_i (y_i - \alpha - \beta x_i^3)(1) = \sum_i y_i - \sum_i \alpha - \beta \sum x_i^3 \Rightarrow \bar{y} - \hat{\alpha} - \hat{\beta}\,\overline{x^3} = 0$$

$$\frac{\partial}{\partial \beta} : \sum_i (y_i - \alpha - \beta x_i^3)\, x_i^3 = \sum y_i x_i^3 - \alpha \sum_i x_i^3 - \beta \sum x_i^6 \Rightarrow \overline{yx^3} - \hat{\alpha}\,\overline{x^3} - \hat{\beta}\,\overline{x^6} = 0$$

hw-R

~3 **17.** Show that SS$_{\text{explained}}$ as defined by $\sum(\hat{y}_i - \bar{y})^2$ can be written as $\hat{\beta}\, S_{xy}$. Hint: use defn of $\hat{y}_i$.

$$SS_{exp} = \sum_i (\hat{y}_i - \bar{y})^2 \underset{\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i}{=} \sum_i (\hat{\alpha} + \hat{\beta} x_i - \bar{y})^2 \underset{\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}}{=} \sum_i (\cancel{\bar{y}} - \hat{\beta}\bar{x} + \hat{\beta} x_i - \cancel{\bar{y}})^2 = \hat{\beta}^2 \underbrace{\sum_i (x_i - \bar{x})^2}_{S_{xx}}$$

$$= \hat{\beta}^2 S_{xx} = \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}} = \hat{\beta}\, S_{xy}.$$

5.54

$\sim \cancel{3}$ 4 **18.** Suppose it is known that the number of wrong bits per minute, x, transmitted over a communication line has the following mass function: $p(x = 0) = 1/2, p(x = 1) = 1/4, p(x = 2) = 1/4$, defined over $x = 0, 1, 2$ only. Also it is known that x is random (technically, independent) from minute to minute. What are the mean and standard deviation of the sampling distribution of the mean of x in one random hour? Show work, but you may leave your answers as fractions.

$$\mu_x = \sum x\,p(x) = 0\left(\tfrac{1}{2}\right) + 1\left(\tfrac{1}{4}\right) + 2\left(\tfrac{1}{4}\right) \Rightarrow \mu_x = \frac{3}{4}$$

$$\sigma_x^2 = \sum (x-\mu_x)^2\,p(x) = \left[\left(0-\tfrac{3}{4}\right)^2\tfrac{1}{2} + \left(1-\tfrac{3}{4}\right)^2\tfrac{1}{4} + \left(2-\tfrac{3}{4}\right)^2\tfrac{1}{4}\right]$$

$$= \frac{1}{4^2}\left(\frac{9}{2} + \frac{1}{4} + \frac{25}{4}\right) = \frac{1}{4^2}11 \Rightarrow \sigma_x = \frac{\sqrt{11}}{4}$$

$$\mu_{\bar{x}} = \mu_x = \frac{3}{4}$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \overset{n=60}{=} \frac{1}{\sqrt{60}} \cdot \frac{\sqrt{11}}{4} = \frac{1}{4}\sqrt{\frac{11}{60}}$$

5.45

~ 2 **19.** Suppose we are interested in MIN(n), the smallest element in a sample of size n, taken from an exponential distribution (with parameter lambda). Will the variance of the sampling distribution of MIN(2) be narrow, comparable, or larger than that of MIN(100)? Explain your reasoning.

Larger

With larger samples (e.g. 100), the min of the sample is more likely to be closer to the true min of the pop. (here 0, because the pop. is exponential). As such, most of the sample mins will be close to zero, hence smaller var. when n=100. So, n=2 var. will be larger than n=100.