Name: ______________________
ID: ______________________
Quiz section or time: ______________________

Stat/Math 390, Winter, Test 2, Feb. 12, 2015; Marzban

AS BEFORE

Points

1 **1.** CO_2 concentration is measured in 100 objects, using a well-established method. A competing method claims to be measuring CO_2 concentration, but more economically. So, the competing method is applied to the **same** 100 objects. Suppose you are interested in knowing whether the two methods are indeed measuring roughly the same quantity. The **best** tool is
a) Histogram of CO_2 concentrations. b) Comparative boxplot of CO_2 concentrations.
c) Scatterplot of CO_2 concentrations. d) QQplot of CO_2 concentrations.

a, b, d all compare histograms, not relationships between 2 things.

1 **2.** Suppose you have found that the best fit to a data set is given by a regression equation of the form $\log(y) = \alpha + \beta(1/x)$. Then, on the average
a) a change of 1 unit in x, leads to a change of β units in y
b) a change of 1 unit in $1/x$, leads to a change of β units in y
c) a change of 1 unit in x, leads to a change of β units in $\log(y)$
d) a change of 1 unit in $1/x$, leads to a change of β units in $\log(y)$

If $\frac{1}{x} \to \frac{1}{x} + 1$
Then $\log y \to \log y + \beta$

1 **3.** It is true that if the relationship between two variables (x, y) is nonlinear and monotonic, then one can transform the data to prepare for simple linear regression modeling. Then, performance of the model based on the transformed data
a) can be assessed with R^2, generally.
b) cannot be assessed with R^2, because of the original nonlinearity.
c) cannot be assessed with R^2, because of the monotonicity.
d) can be assessed with R^2, only if there is no interaction term.

$R^2 = SS_{expl}/SST$ for any model, any data

4. Circle all the correct statements. The correlation coefficient
1 a) is misleading if the scatterplot contains cluster or outliers.
b) is misleading for nonlinearly related data.
c) measures the slope of a line going through the scatterplot. skinniness not slope
d) is a measure of skinines about the OLS (best fit) line. r does not presume a fit at all.

1 **5.** Consider the OLS model: $y = 1 + 2x_1 + 2x_2 + \boxed{4x_1x_2}$. Which of the following is/are true?
a) The average rate of change of y with respect to 1 unit change in x_1, if x_2 is held constant, is 2.
b) There is evidence for collinearity c) The average value of y is 1 d) None of the above

1 **6.** Suppose x has a Poisson distribution with parameter λ (i.e., its mean and variance are both λ). The expected value and variance of the **sampling distribution** of the sample mean are
a) $\overline{x}, s_x^2$ b) λ, λ c) $\lambda, s_x^2/n$ d) $\lambda, \lambda/n$ e) Cannot tell, because population is not normal.

$E[\bar{x}] = \mu_x = \lambda$ $V[\bar{x}] = \sigma_x^2/n = \lambda/n$

1 **7.** For which of the following quantities does a sampling distribution NOT exist?
a) Population mean b) Sample variance c) Sample minimum d) All of the above.

~ 2 **8.** The qqplot for the following situations is expect to be (approximately) a straight line; specify the y-intercept and slope.

	y-int	slope
a) Data are from $N(0,1)$, and the x-axis corresponds to quantiles of $N(0,1)$	0	1
b) Data are from $N(2,3)$, and the x-axis corresponds to quantiles of $N(0,1)$	2	3
c) Data are from $N(2,3)$, and the x-axis corresponds to quantiles of $N(2,3)$	0	1
d) Data are from Exp$(\lambda = 1)$, and the x-axis corresponds to quantiles of Exp$(\lambda = 1)$	0	1

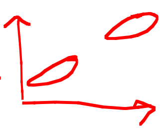
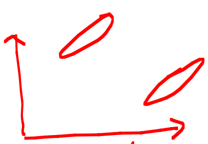
~ 2 **9.** In a regression model of the type $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$, briefly explain the argument for why the β_i are not interpretable (as average rate of change of y w.r.t. a unit change in x_i).

Answer 1: Because $y = \alpha + \beta_1 x_1 + \underbrace{(\beta_2 + \beta_3 x_1)}\, x_2$, i.e. the rate of change of y w.r.t. x_2 depends on x_1. Same for x_1.

Answer 2: rate of change of y w.r.t. x_2: $\partial y / \partial x_2 = \beta_2 + \beta_3 x_1$ depends on x_1. Same for x_1.

~ 2 **10.** Suppose the correlation coefficient r_{xy} between two variables x and y is positive when for that data set some other variable w is 0. Suppose r_{xy} is also positive in some other data set where w is some other value (say 1). What can one say about the sign of r_{xy} in the combined data set? Explain in words and/or by figures.

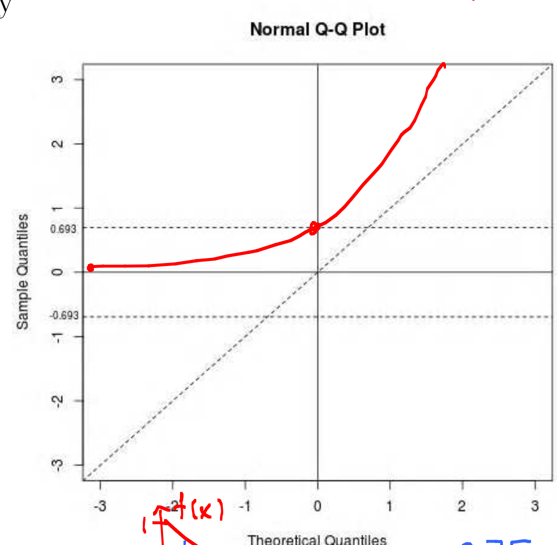
Not much. The r in the combined data set may be

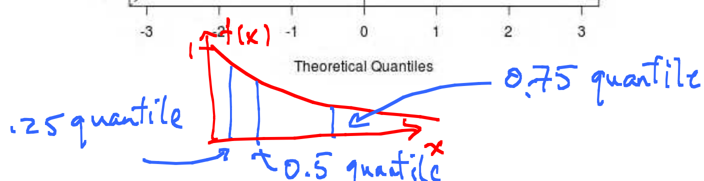
positive:  or zero:  or negative: 

(simpson's paradox)

~ ~~2~~ 2.5 **11.** On the adjacent plot, draw the qqplot as accurately as you can, for a situation where data come from an exponential dist. with param. $\lambda = 1$, and the xaxis corresponds to the quantiles of standard Normal. Note: the median of Exponential($\lambda = 1$) is $ln(2) = 0.693$. SHOW WORK!

	Low quantile	0.5 quantile	high quantile
Std. Normal	-3	0	3
Exp(1)	0	.693	large





~ 2 **12.** In simple linear regression model of the form $y = \alpha + \beta x$, where α and β have been estimated from a data set for which $\bar{x}, \bar{y}, \overline{xy}, \overline{x^2}, \overline{y^2}$ are all known, find the predicted value of y (in terms of the known quantities) when $x = \bar{x}$? Show work.

At $x = \bar{x}$: $\hat{y} = \hat{\alpha} + \hat{\beta}\bar{x} = \bar{y} - \cancel{\hat{\beta}\bar{x}} + \cancel{\hat{\beta}\bar{x}} = \bar{y}$.

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

In English: The OLS fit goes through the point corresponding to the sample means of x, and y: $(\bar{x}, \bar{y})$.

$\sim \cancel{2}$ 2.5 **13.** In a multiple regression problem, the model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$ is employed. The data contains 15 cases, the sample standard deviation of the y is 21, and the typical error/deviation of the data from the fitted surface is 7. Compute R^2 (in terms of numbers, not symbols, but don't waste time on arithmetic.)

$$R^2 = \frac{SS_{expl}}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{[n-(k+1)]\,S_e^2}{(n-1)\,S_y^2} \qquad S_e^2 = \frac{SSE}{n-(k+1)}, \quad S_y^2 = \frac{SST}{(n-1)}$$

$k = 5$

$$= 1 - \frac{15-(5+1)}{15-1} \cdot \left(\frac{7}{21}\right)^2 = 1 - \frac{9}{14}\cdot\frac{1}{9} = \boxed{\frac{13}{14}}$$

S_e, S_y

$\sim \cancel{2}$ 3 **14.** Consider the OLS estimate of the slope parameter in $y = \alpha + \beta x$. Derive the relationship between that estimate and the estimate one would get if x and y are switched. Show work.

Way 1: $\hat\beta = \dfrac{\overline{xy} - \bar x \bar y}{\overline{x^2} - \bar x^2} \rightarrow \dfrac{\overline{yx} - \bar y \bar x}{\overline{y^2} - \bar y^2} = \dfrac{\overline{xy} - \bar x \bar y}{\overline{y^2} - \bar y^2} = \dfrac{\overline{xy} - \bar x \bar y}{\overline{y^2} - \bar y^2} \cdot \dfrac{\overline{x^2} - \bar x^2}{\overline{x^2} - \bar x^2} = \hat\beta \left(\dfrac{\overline{x^2} - \bar x^2}{\overline{y^2} - \bar y^2}\right)$

Way 2: $\hat\beta = \dfrac{S_{xy}}{S_{xx}} \longrightarrow \dfrac{S_{yx}}{S_{yy}} = \dfrac{\sum_i (x_i - \bar x)(y_i - \bar y)}{S_{yy}} = \dfrac{S_{xy}}{S_{yy}} \cdot \dfrac{S_{xx}}{S_{xx}} = \dfrac{S_{xy}}{S_{xx}} \cdot \dfrac{S_{xx}}{S_{yy}} = \hat\beta \left(\dfrac{S_{xx}}{S_{yy}}\right)$

Way 3: $\hat\beta \cdot \hat\beta' = r^2$

$\sim \cancel{2}$ 2.5 **15.** Find the normal equations for the OLS estimates of α and β for the model $y_i = \alpha + \beta\, log(x_i) + \epsilon_i$. It's important that you do it this way: Start from the expression for SSE, differentiate, and re-write the expressions in terms of "barred" quantities (i.e., averages). No need to solve for the estimates.

$$SSE = \sum_i \epsilon_i^2 = \sum_i \left(y_i - \alpha - \beta \log(x_i)\right)^2$$

$$\frac{\partial}{\partial\alpha}: \sum_i (y_i - \hat\alpha - \hat\beta \log(x_i)) = 0 \implies \bar y - \hat\alpha - \hat\beta\, \overline{\log x} = 0$$

$$\frac{\partial}{\partial\beta}: \sum_i (y_i - \hat\alpha - \hat\beta \log(x_i)) \cdot \log(x_i) = 0 \implies \overline{y\log(x)} - \hat\alpha\, \overline{\log x} - \hat\beta\, \overline{(\log(x))^2} = 0$$

~ 3 **16.** A sampling distribution (e.g. of the sample mean) is a distribution, not a histogram of data (the histogram is just an intuitive way of understanding the distribution). One can derive it mathematically, if one knows the population distribution. Let's do one. Consider a population/distribution described by $p(x = 0) = (1 - \pi)$, $p(x = 1) = \pi$. So, x takes values 0 or 1, and the parameter of the pop is π. Find/derive the sampling distribution of the sample mean, for samples of size 2. Hint: write down the possible samples, the corresponding value of $\bar x$, and the probability for each; remember how we derived the binomial distribution.

Possible $n=2$ samples:	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$\bar x =$	0	$\frac{1}{2}(0+1) = \frac{1}{2}$	$\frac{1}{2}(1+0) = \frac{1}{2}$	$\frac{1}{2}(1+1) = 1$
prob $=$	$(1-\pi)^2$	$(1-\pi)\pi$	$\pi(1-\pi)$	π^2

$p(\bar x = 0) = \text{prob}(\bar x = 0) = (1-\pi)^2$

$p(\bar x = \frac{1}{2}) = \text{prob}(\bar x = \frac{1}{2}) = \pi(1-\pi) + (1-\pi)\pi = 2\pi(1-\pi)$

$p(\bar x = 1) = \text{prob}(\bar x = 1) = \pi^2$