Homework 1

E6690: Statistical Learning for Bio & Info Systems

P1. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Show that:

(a) (2pt) $\sum_{i=1}^{n} X_i^2 = (n-1)S^2 + n\bar{X}^2$

(b) (2pt) If $X_1, X_2, ..., X_n$ are independent and identically distributed (i.i.d.), the S^2 is an unbiased estimator of σ^2, i.e., $\mathbb{E}S^2 = \sigma^2$

In the following, in addition to the above, assume that X_i-s have normal/Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$.

(c) (3pt) Show (prove) that $\bar{X}$ is independent of $X_i - \bar{X}$, $i = 1, 2, \ldots, n$.
(Hint: Both $\bar{X}$ and $X_i - \bar{X}$ are normal.)

(d) (3pt) Show (prove) that the sample mean, $\bar{X}$, is independent of the sample variance, S^2.

P2. (10pt) Show that in the case of simple linear regression between Y and X, the R^2 statistic is equal to the square of the correlation coefficient between X and Y (r^2). For simplicity, you may assume that $\bar{y} = \bar{x} = 0$. Recall that

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{and} \quad r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

P3. (20pt; each bullet 2pt) Create some simulated data and fit simple linear regression models to it. Make sure to use `set.seed(1)` prior to starting part (a) to ensure consistent results.

(a) Using the `rnorm()` function, create a vector, `x`, containing 100 observations drawn from a $\mathcal{N}(0, 1)$ distribution. This represents a feature, X.

(b) Using the `rnorm()` function, create a vector, `eps`, containing 100 observations drawn from a $\mathcal{N}(0, 0.25)$ distribution.

(c) Using `x` and `eps`, generate a vector `y` according to the model

$$Y = -1 + 0.5X + \epsilon.$$

What is the length of the vector `y`? What are the values of β_0 and β_1 in this linear model?

(d) Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe.

(e) Fit a least squares linear model to predict `y` using `x`. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

(f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the `legend()` command to create an appropriate legend.

(g) Now fit a polynomial regression model that predicts `y` using `x` and `x`^{2}. Is there evidence that the quadratic term improves the model fit? Explain your answer.

(h) Repeat (a)-(f) after modifying the data generation process in such a way that there is less noise in the data. The model in (c) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.

(i) Repeat $(a) - (f)$ after modifying the data generation process in such a way that there is more noise in the data. The model in (c) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.

(j) What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results. (You could use the `confint()` function.)

P4. (10pt) Using R and `Advertising` data set, find 92% confidence intervals for β_0 and β_1 for three single-feature linear regressions of `Sales` versus `Newspaper`, `TV` and `Radio`, respectively. Then, create a scatterplot for each of them with the 92% confidence interval lines, i.e., draw the lines that correspond to the ends of confidence intervals for (β_0, β_1). The answer should include the R code and graphs.

P5. Consider the `Auto` data set:

(a) (5pt) Produce a scatterplot matrix which includes all of the pairs of variables in the data set.

(b) (5pt) Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the `name` variable, which is qualitative.

(c) (5pt) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

i. Is there a relationship between the predictors and the response?

ii. Which predictors appear to have a statistically significant relationship to the response?

iii. What does the coefficient for the `year` variable suggest?

(d) (5pt) Try a few different transformations of the variables, such as $\log(X)$, $\sqrt{X}$, X^2. Comment on your findings.

P6. (10pt) A data set has $n = 20$,

$$\sum_{i=1}^{20} x_i = 8.552, \quad \sum_{i=1}^{20} y_i = 398.2, \quad \sum_{i=1}^{20} x_i^2 = 5.196, \quad \sum_{i=1}^{20} y_i^2 = 9356, \quad \text{and} \quad \sum_{i=1}^{20} x_i y_i = 216.6.$$

Calculate $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$. What is the fitted value when $x = 0.5$? Compute R^2.

P7. (10pt) The multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

is fitted to a data set of $n = 45$ observations. The total sum of squares is TSS $= 11.62$, and the residual sum of squares is RSS $= 8.95$. What is the p-value for the null hypothesis

$$\mathcal{H}_0 : \quad \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \quad ?$$

Extra Credit

Under normal assumptions we can compute the distributions of a lot of quantities explicitly.

E1. (5pt) *Chi-squared distribution.* Let $X_1, X_2, \ldots, X_n$ be independent standard normal random variables and recall that Chi-squared random variable with n degrees of freedom is defined as $\chi^2_n = X_1^2 + X_2^2 + \cdots + X_n^2$. Prove that the density of χ^2_n is given by

$$g_n(x) = \frac{1}{\Gamma(n/2)2^{n/2}} x^{n/2-1} e^{-x/2},$$

where $\Gamma(x)$ is the gamma function. (Hint: Prove first for $n = 1, 2$, and then use the mathematical induction.)

E2. (5pt) Let $X_1, X_2, \ldots, X_n$ be independent normal random variables $\mathcal{N}(\mu, \sigma^2)$. Prove that

$$\frac{(n-1)S^2}{\sigma^2} \stackrel{d}{=} \chi^2_{n-1},$$

where $\stackrel{d}{=}$ stands for equality in distribution.
(Hint: Derive the moment generating function of χ^2_n and use problem **P1.**(a) and (d).)

E3. (5pt) *Student's t distribution.* Let t_n be student's t variable, defined as

$$t_n = \frac{Z}{\sqrt{\chi^2_n/n}},$$

where $Z \sim \mathcal{N}(0, 1)$. Prove that t_n has the density

$$f_n(t) = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\Gamma(n/2)} \cdot \frac{1}{(1 + t^2/n)^{(n+1)/2}},$$

where $\Gamma(x)$ is the gamma function. Show that for large values of n, $f_n(t)$ is approximately normal, $f_n(t) \approx e^{-t^2}/\sqrt{2\pi}$. (Hint: First show that the conditional density (distribution) of t_n given $\chi^2_n = x$ is normal with mean 0 and variance $\sqrt{n/x}$. Then, use problem **E1.** to integrate this conditional density.)

E4. (5pt) *F (Fisher) distribution.* Let U and V be two independent Chi-squared random variables with degrees of freedom n_1 and n_2, and define the random variable, $F \equiv F(n_1, n_2)$, as

$$F = \frac{U/n_1}{V/n_2}.$$

Show that the density of F is given by

$$f_{n_1,n_2}(w) = \frac{(n_1/n_2)^{n_1/2}\Gamma[(n_1+n_2)/2]w^{(n_1/2)-1}}{\Gamma[n_1/2]\Gamma[n_2/2][1 + (n_1 w/n_2)]^{(n_1+n_2)/2}}.$$

(Hint: Compute first the distribution of F given V.)