STA 6207 – Homework 2

Q.1. Model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1,\ldots,n$

p.1.a. Derive the normal equations that minimize $Q = \sum_{i=1}^{n} \varepsilon_i^2$.

p.1.b. Solve for the ordinary least squares estimators $\hat{\beta}_1, \quad \hat{\beta}_0$

p.1.c. Derive $E\left\{\hat{\beta}_1\right\}, \quad V\left\{\hat{\beta}_1\right\}, \quad E\left\{\hat{\beta}_0\right\}, \quad V\left\{\hat{\beta}_0\right\}, \quad \mathrm{COV}\left\{\hat{\beta}_0, \hat{\beta}_1\right\}$

p.1.d. Derive the mean and variance of $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ and $e_i = Y_i - \hat{Y}_i$ and $\mathrm{COV}\left\{\hat{Y}_i, e_i\right\}$

Q.2. An electrical contractor fits a simple linear regression model, relating cost to wire a house (Y, in dollars) to the size of the house (X, in ft^2). She fits a model, based on a sample of n=16 houses and obtains the following results.

$$\hat{Y} = 50.00 + 0.22X \qquad s^2 = 1600 \qquad \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = 4000000 \qquad \overline{X} = 2000$$

p.2.a. Compute the estimated standard errors of $\hat{\beta}_1$ and $\hat{\beta}_0$

p.2.b. Compute a 95% Confidence Interval for β_1

p.2.c. Compute a 95% Confidence Interval for the mean of all homes with $X_0 = 2000$

p.2.d. Compute a 95% Prediction Interval for her brother-in-laws house with $X_0 = 2000$

Q.3. A researcher is interested in the relationship between the education level and salaries in rural counties in the U.S. He obtains the percentage of adults over 25 with a college education in each county (X) and the per capita income of the county (Y, in $1000s). He obtains the following summary statistics, based on a sample of n= 30 counties.

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = 2207.45 \quad \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = 658.37 \quad \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = 654.86 \quad \overline{X} = 41.92 \quad \overline{Y} = 35.83$$

p.3.a. Compute least squares estimates of β_0 and β_1

p.3.b. Show that $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 - \dfrac{\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right]^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$

p.3.c. Use p.3.b. to compute an unbiased estimate of σ^2

Q.4. Consider the regression through the origin model: $Y_i = \beta_1 X_i + \varepsilon_i \quad \varepsilon_i \sim NID\left(0, \sigma^2\right) \quad i = 1, \ldots, n$

p.4.a. Derive the least squares estimator of β_1

p.4.b. Derive the mean and variance of the least squres estimator.

p.4.c. Consider the estimator $\tilde{\beta}_1 = \dfrac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}$. Derive its mean and variance.

p.4.d. Which estimator has the smallest variance? Why?

Q.5. For the simple linear regression model with an intercept, show that $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right) = 0$

Q.6. For simple regression, we get: $\hat{\beta}_1 = \sum_{i=1}^{n}\left(\dfrac{X_i - \overline{X}}{S_{XX}}\right) Y_i$ and $\overline{Y} = \sum_{i=1}^{n}\left(\dfrac{1}{n}\right) Y_i \qquad \text{COV}\left(\hat{\beta}_1, \overline{Y}\right) = ???$

Q.7. For a simple linear regression model, derive $\text{COV}\left\{\hat{\beta}_0, \hat{\beta}_1\right\}$ completing the following parts:

p.7.a. Write $\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$ and $\hat{\beta}_0 = \sum_{i=1}^{n} b_i Y_i$ stating explicitly what the a_i and b_i values (functions) are

p.7.b. Using rules of Covariances of linear functions of random variables to derive $\text{COV}\left\{\hat{\beta}_0, \hat{\beta}_1\right\}$

p.7.c. Researchers in the U.S. fit regressions of relationship between viscosity (Y) and temperature (X) in degrees Fahrenheit, while foreign researchers work with temperature in degrees Celsius. The temperatures for the experimental runs are given below. Give the $\text{COV}\left\{\hat{\beta}_0, \hat{\beta}_1\right\}$ for each set of researchers as a function of σ^2 (they use the same units for Y):

Run #	1	2	3	4	5	6
X(F)	5	5	14	14	23	23
X(C)	-15	-15	-10	-10	-5	-5

$\text{COV}\left\{\hat{\beta}_0, \hat{\beta}_1\right\}$ Fahrenheit = ____________________________ Celsius = ______________________________

Q.8. An Austrian study considered Breath Alcohol Elimination Rates (X, mg/L/hr*100) and Blood Alcohol Elimination Rates (Y, g/L/hr*100) in a sample of n = 27 adult females. The sample means, standard deviations and correlations are given below. Complete the following table for the simple linear regression relating Blood Elimination Rate (Y) to Breath Elimination Rate (X).

$$\overline{X} = 8.6704 \quad \overline{Y} = 17.8815 \quad s_X = 1.6522 \quad s_Y = 3.6787 \quad r_{XY} = 0.8786$$

Regression Statistics					
R Square					
Residual Std Error					
Observations					
ANOVA					
	df	*SS*	*MS*	*F*	*F(.05)*
Regression					
Residual					
Total					
	Coefficients	*andard Err*	*t Stat*	*t(.025)*	
Intercept					
X					

Q.9. For the simple regression model (scalar form): $Y_i = \beta_0 + \beta_1\left(X_i - \overline{X}\right) + \varepsilon_i \quad i = 1,...,n \quad \varepsilon_i \sim NID\left(0, \sigma^2\right)$

p.9.a. Derive least squares estimators of β_0 and β_1.

p.9.b. Derive $E\left\{\hat{\beta}_1\right\}, \quad E\left\{\hat{\beta}_0\right\}, \quad V\left\{\hat{\beta}_1\right\}, \quad V\left\{\hat{\beta}_0\right\}, \quad \mathrm{COV}\left\{\hat{\beta}_0, \hat{\beta}_1\right\}$

Q.10. Consider the "centered" (with respect to the independent variable) model in scalar form:

$$Y_i = \mu + \beta_1\left(X_i - \overline{X}\right) + \varepsilon_i \quad i = 1,...,n \quad \varepsilon_i \sim NID\left(0, \sigma^2\right)$$

p.10.a. Obtain the normal equations and the least squares estimated for the parameters μ and β_1.

p.10.b. Derive $COV\left(\hat{\mu}, \hat{\beta}_1\right)$ Hint: $COV\left(\sum_{i=1}^{n} a_i Y_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j COV\left(Y_i, Y_j\right)$

Q.11. A linear regression was run on a set of data, based on a simple linear regression. You are given only the following partial information:

ANOVA						
	df	*SS*	*MS*	*F*	*P-value*	
Regression						
Residual	5		44.2			
Total						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*		
Intercept	293.89	5.62	52.29	0.0000		
X	-1.65	0.13	-13.13	0.0000		

p.11.a. Compute a 95% Confidence Interval for β_1:

p.11.b. Give the F-statistic and rejection for testing H_0: $\beta_1 = 0$ vs H_A: $\beta_1 \neq 0$ at $\alpha = 0.05$ significance level. (Hint: think of connection between t- and F-tests)

p.11.c. Compute the coefficient of determination, R^2.

Q.12. A regression model was fit, relating revenues (Y) to total cost of production and distribution (X) for a random sample of n=30 RKO films from the 1930s (the total cost ranged from 79 to 1530):

$$n = 30 \quad \overline{X} = 685.2 \quad S_{xx} = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = 6126371 \quad \hat{Y} = 55.23 + 0.92X \quad S_e^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2} = 40067$$

p.12.a. Obtain a 95% Confidence Interval for the **mean revenues for all movies** with total costs of $x^* = 1000$

Note: $$\left[\frac{1}{30} + \frac{(1000-685.2)^2}{6126371}\right] = 0.0495$$

$\hat{\mu}_y =$ ____________ $SE_{\hat{\mu}} =$ ______________ $95\%\text{CI}:$ ____

p.12.b. Obtain a 95% Prediction Interval for **tomorrow's new film** that had total costs of $x^* = 1000$

$\hat{y} =$ ____________ $SE_{\hat{y}} =$ ______________ $95\%\text{PI}:$ ____

Q.13.

```
Dataset:  std_intensity.dat

Source: T.R.M. De Beer, G.J. Vergote, W.R.G. Baeyens, J.P. Remon, C. Vervaetand F. Verpoort (2004).
"Development and Validation of a Direct, non-Destructive Quantitative Method for
MedroxyprogesteroneAcetate in a Pharmaceutical Suspension Using FT-Raman Spectroscopy," European
Journalof Pharmaceutical Sciences, Vol. 23, pp. 355-362

Description: Simple Linear Regression relating measured peak intensities (Y)
to standard suspension Concentration. 4 concentrations, 6 reps/conc.

Variables/Columns
Standard Suspension Concentration   (mg/ml)  1-8
Measured Peak Intensities  (AU)   10-16
```

Fit a simple linear regression model, relating measured peak intensity (Y) to standard suspension concentration (X) using "brute-force" computations and the lm function in R. Give the following results (clearly stating all elements of tests).

a) Sample size, sample means, sums of squares and sum of crossproducts
b) Least squares estimates of $\beta_1$and β_0 and unbiased estimate of σ^2
c) Estimated standard errors, t-statistics, P-values, and 95% CI's regarding β_1 and β_0
d) Total SS, Error SS, Regression SS, Analysis of Variance, and F-test
e) 95% Confidence Interval for Mean and 95% Prediction Interval for single measurement when X = 140.

Q.14. Oddsmakers predictions of total scores in Women's NBA 2010-2018 regular season games (with overtime games removed). The following program obtains the population model among all games from the 9 seasons.

```
wnba1 <- read.csv("http://users.stat.ufl.edu/~winner/data/wnba_spread.csv")
attach(wnba1); names(wnba1)

# Model for the Population of Games
Y <- totPts[OT == 0];    X <- OU[OT == 0]
(N <- length(Y))
(rho <- cor(X,Y))
(beta1 <- cov(X,Y) / var(X))
(beta0 <- mean(Y) - mean(X) * beta1)
E.Y <- beta0 + beta1 * X
eps <- Y - E.Y
(sigma2 <- sum(eps^2) / N)
summary(eps)

win.graph(height=5.5, width=7.0)
plot(Y ~ X, pch=16, cex=0.5)
abline(lm(Y~X), col="red", lwd=3)

ymax <- 1.3*2*N*dnorm(0, 0 , sqrt(sigma2))
hist(eps, breaks=seq(-50,50,2), xlab="eps", ylim=c(0, ymax))
lines(seq(-50,50,.1), N*2*dnorm(seq(-50,50,.1),0,sqrt(sigma2)))
```

Generate 10000 random samples of size n = 25, and for each sample obtain 95% Condidence Intervals for

β_0, β_1, ρ, σ^2 ($SSE/\sigma^2 \sim \chi_{n-2}$). You will also need to save: $\hat{SE}\left\{\hat{\beta}_0\right\}, \quad \hat{SE}\left\{\hat{\beta}_1\right\}$

a) Obtain the empirical coverage rates of the 95% CI's for β_0, β_1, ρ, σ^2
b) Obtain histograms of : $\dfrac{\hat{\beta}_0-\beta_0}{\hat{SE}\left\{\hat{\beta}_0\right\}}, \quad \dfrac{\hat{\beta}_1-\beta_1}{\hat{SE}\left\{\hat{\beta}_1\right\}}, \quad \dfrac{SSE}{\sigma^2}=\dfrac{(n-2)MSE}{\sigma^2}$