Simple Linear Regression - Scalar Form

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 $E\left\{\hat{\beta}_1\right\}, \quad V\left\{\hat{\beta}_1\right\}, \quad E\left\{\hat{\beta}_0\right\}, \quad V\left\{\hat{\beta}_0\right\}, \quad \text{COV}\left\{\hat{\beta}_0, \hat{\beta}_1\right\}$

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 $\text{COV}\left\{\hat{Y}_i, e_i\right\}$

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 $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 - \frac{\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right]^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$

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

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

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 $\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}$. 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: $\hat{\beta}_1 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{S_{XX}} \right) Y_i$ and $\overline{Y} = \sum_{i=1}^{n} \left(\frac{1}{n} \right) Y_i$ $\quad$ $\text{COV}\left(\hat{\beta}_1, \overline{Y} \right) = ???$

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

we get: $\hat{\beta}_1 = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{S_{XX}}\right)Y_i, \quad \overline{Y} = \sum_{i=1}^{n}\left(\frac{1}{n}\right)Y_i, \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

p.8.a. Derive $E\left(\hat{\beta}_1\right), \quad E\left(\overline{Y}\right), \quad E\left(\hat{\beta}_0\right)$

p.8.b. Derive $V\left(\hat{\beta}_1\right), \quad V\left(\overline{Y}\right), \quad \mathrm{COV}\left(\hat{\beta}_1, \overline{Y}\right), \quad V\left(\hat{\beta}_0\right), \quad \mathrm{COV}\left(\hat{\beta}_1, \hat{\beta}_0\right)$

Q.9. 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.10. 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.10.a. Derive least squares estimators of β_0 and β_1.

p.10.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.11. Consider the "centered" (with respect to the independent variable) model in scalar form:

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

p.11.a. Obtain the normal equations and the least squares estimated for the parameters μ and β_1.

p.11.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.12. 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.12.a. Compute a 95% Confidence Interval for β_1:

p.12.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.12.c. Compute the coefficient of determination, R^2.

Q.13. A simple linear regression model is to be fit, relating Mean Annual Temperature (Y, in °F) to Year – 1957 (X). That is, the "origin" is 1957. The data are for the years 1957-2014 (n = 58). The sample means and sums of squares and cross-products are given below for Model 1.

$$\overline{X} = 28.5 \quad \overline{Y} = 68.2759 \quad \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = 559.6083 \quad \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = 16254.5 \quad \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = 83.4361$$

p.13.a. Compute the least squares estimates of β_1 and β_0, and write out the predicted equation.

p.13.b. $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = 64.1699$ Use this to obtain an unbiased estimate of σ^2.

p.13.c. Obtain a 95% Confidence Interval for β_1 (the amount, on average, that mean temperature increases per year).

p.13.d. For the year 2014, the (observed) average temperature was 67.65. Obtain the predicted temperature and the residual for that year.

Q.14. Derive the two normal equations by minimizing $Q = \sum_{i=1}^{n}\varepsilon_i^2$ with respect to β_0 and β_1.

Q.15. The fitted value for Y_j can be written as $\hat{Y}_j = \sum_{i=1}^{n}\left[\frac{1}{n} + \frac{\left(X_i - \overline{X}\right)}{SS_{XX}}\left(X_j - \overline{X}\right)\right]Y_i$

p.15.a. Give $\text{COV}\left\{Y_j, \hat{Y}_j\right\}$

p.15.b. Give $\text{COV}\left\{Y_k, \hat{Y}_j\right\}$ $k \neq j$

Q.16. Derive $E\{MSR\}$ and $E\{MSE\}$

Q.17. A simple linear regression model is fit relating fairway accuracy (Y, in percent) to average drive distance (X, in yards) for a random sample of n = 20 professional women golfers from the 2009 season.

p.17.a. Complete the following table cells.

Regression Statistics					
R Square					
Observations	20				
ANOVA					
Source	*df*	*SS*	*MS*	*F*	*F(.05)*
Regression	1	81.7	81.7		
Residual				#N/A	#N/A
Total		407.5	#N/A	#N/A	#N/A
Parameter	*Coefficients*	*Standard Error*	*t Stat*	*t(.025)*	
Intercept	131.517	29.552	4.450		
drive	-0.250	0.118			

p.17.b. What is your conclusion regarding the test of $H_0: \beta_1 = 0 \quad H_A: \beta_1 \neq 0$? **Reject H_0 / Fail to Reject H_0**

p.17.c. Give the sample correlation, *r*, between Y and X.

Q.18. A simple linear regression model is fit, relating plant growth over 1 year (y) to amount of fertilizer provided (x). Twenty five plants are selected, 5 each assigned to each of the fertilizer levels (12, 15, 18, 21, 24). The results of the model fit are given below:

Coefficients[a]

Model		Unstandardized Coefficients				
		B	Std. Error		t	Sig.
1	(Constant)	8.624	1.810		4.764	.000
	x	.527	.098		5.386	.000

a. Dependent Variable: y

p.18.a. Can we conclude that there is an association between fertilizer and plant growth at the 0.05 significance level? Why (be very specific).

p.18.b. Give the estimated mean growth among plants receiving 20 units of fertilizer.

p.18.c. The estimated standard error of the estimated mean at 20 units is $2.1\sqrt{\frac{1}{25} + \frac{(20-18)^2}{450}} = 0.46$

Give a 95% CI for the mean at 20 units of fertilizer.

Q.19. A study was conducted to relate weight gain in chickens (Y) to the amount of the amino acid lysine ingested by the chicken (X). A simple linear regression is fit to the data.

ANOVA					
	df	SS	MS	F	P-value
Regression	1	27.07	27.07	23.79	0.0012
Residual	8	9.10	1.14		
Total	9	36.18			

	Coefficients	Standard Error	t Stat	P-value
Intercept	12.4802	1.2637	9.8762	0.0000
lysine(X)	36.8929	7.5640	4.8774	0.0012

p.19.a. Give the fitted equation, and the predicted value for X=0.20
p.19.b. Give a 95% Confidence Interval for the MEAN weight gain of all chickens with X=0.20 (Note: the mean of X is 0.16 and S_{XX}=0.020)
p.19.c. What proportion of the variation in weight gain is "explained" by lysine intake?

Q.20. A researcher reports that the correlation between length (inches) and weight (pounds) of a sample of 16 male adults of a species is r=0.40.

p.20.a. Test whether she can conclude there is a POSITIVE correlation in the population of all adult males of this species:

$H_0: \rho = 0$ $\quad H_A: \rho > 0$

- Test Statistic:
- Rejection Region (α=0.05):
- Conclude: **Positive Association** or **No Positive Association**

p.20.b. A colleague from Europe transforms the data from length in inches to centimeters (1 inch=2.54 cm) and weight from pounds to kilograms (1 pound=2.2 kg). What is the colleague's estimate of the correlation?

Q.21. A simple linear regression is to be fit, relating fuel efficiency (*Y* in gallons/100 miles) to cars weight (*X*, in pounds), based on a sample of *n*=45 cars. You are given the following information:

$$\sum_{i=1}^{n}(X_i-\overline{X})^2=13069326 \quad \sum_{i=1}^{n}\left(X_i-\overline{X}\right)(Y_i-\overline{Y})=13385 \quad \sum_{i=1}^{n}(Y_i-\overline{Y})^2=16.5$$

$$\overline{X}=2739 \quad \overline{Y}=3.4 \quad \sum\left(Y-Y\right)^2=2.835$$

Compute the following quantities:

p.21.a. β_1 ____

p.21.b. β_0 ____

p.21.c. Residual Std. Deviation s_e ________________________________

p.21.d. Estimate of mean efficiency for all cars of x*=2000 pounds ________________________________

p.21.e. 95% Confidence Interval for all cars of *x**=2000 pounds

Lower Bound = ________________________ Upper Bound = ___________________________

p.21.f. Regression Sum of Squares *SSR* = ____

p.21.g. Proportion of Variation in Efficiency "Explained" by Weight _________________________

Q.22. 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.22.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%CI: ________________________________

p.22.b. Obtain a 95% Prediction Interval for **tomorrow's new film** that had total costs of $x^* = 1000$

$\hat{y} =$ __________ $SE_{\hat{y}} =$ ___________ 95%PI: ________________________________

Q.23. A researcher is interested in the correlation between height (*X*) and weight (*Y*) among 12 year old male children. He selects a random sample of *n* = 18 male 12-year olds from a school district, and intends on testing $H_0: \rho = 0$ versus $H_A: \rho \neq 0$, where ρ is the population correlation coefficient.

His sample correlation is $r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} = 0.60$

Test $H_0: \rho = 0$ versus $H_A: \rho \neq 0$:

Test Statistic = ________________________ Rejection Region: __________________________

Q.24. A study was conducted to determine the effects of daily temperature (X, in °C) on Electricity Consumption (Y, in 1000s of Wh) in an experimental house over a period of n = 31 days. Consider the following model:

$$E\{Y\} = \beta_0 + \beta_1 X \quad SSR = 594.0 \quad SSE = 241.4 \quad TSS = 835.4 \quad S_{XX} = 158.5 \quad \overline{X} = 27.0 \quad \hat{Y} = -30.179 + 1.936X$$

p.24.a. What proportion of the variation in Electricity consumption is "explained" by daily temperature (X)?

p.24.b. Compute the residual standard deviation, s_e

p.24.c. Obtain the estimated **mean** electricity consumption when $x^* = 27.0$ degrees, and the 95% Confidence Interval.

Estimated Mean: ________________________________ 95% CI: __________________________________

p.24.d. Compute a 95% Confidence Interval for ρ.

Q.25. For the simple regression model (scalar form): $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1,...,n \quad \varepsilon_i \sim NID\left(0,\sigma^2\right)$

we get: $\hat{\beta}_1 = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{S_{XX}}\right)Y_i, \quad \overline{Y} = \sum_{i=1}^{n}\left(\frac{1}{n}\right)Y_i, \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

p.25.a. Derive $E\left(\hat{\beta}_1\right), \quad E\left(\overline{Y}\right), \quad E\left(\hat{\beta}_0\right)$

p.25.b. Derive $V\left(\hat{\beta}_1\right), \quad V\left(\overline{Y}\right), \quad \text{COV}\left(\hat{\beta}_1, \overline{Y}\right), \quad V\left(\hat{\beta}_0\right), \quad \text{COV}\left(\hat{\beta}_1, \hat{\beta}_0\right)$

Q.26. 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.26.a. Derive least squares estimators of β_0 and β_1.

p.26.b. Derive $E\left\{\hat{\beta}_1\right\}, \quad E\left\{\hat{\beta}_0\right\}, \quad V\left\{\hat{\beta}_1\right\}, \quad V\left\{\hat{\beta}_0\right\}, \quad \text{COV}\left\{\hat{\beta}_0, \hat{\beta}_1\right\}$

Q.27. An experiment was conducted to relate dry matter digestibility (Y, in percent) to concentration of neutral detergent fiber (X) in horse diets. The authors reported the following results.

$n = 65 \quad \overline{X} = 61.4 \quad s_X = 13.71 \quad \overline{Y} = 52.7 \quad s_Y = 8.87 \quad \hat{Y} = 85.42 - 0.5324X \quad \sqrt{MSE} = 5.966 \quad r^2 = 0.6017$

p.27.a. Obtain $\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2, \quad \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \quad \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$

p.27.b. Compute the 95% Confidence Interval for mean dry matter digestibility neutral detergent fiber has a concentration of 60.

p.27.c. Complete the Following table.

ANOVA						
	df	*SS*	*MS*	*F*	*F(.05)*	
Regression						
Residual						
Total						
	Coefficients	*Standard Err*	*t Stat*	*t(.025)*	*Lower 95%*	*Upper 95%*
Intercept						
X						

Q.28. Based on model 1, derive the ordinary least squares estimators of β_1, β_0. **Show all work.**

Q.29. A study in Beijing, China related heights (Y, cm) to mean foot length (X, cm) in a sample of n = 174 male college students. The sample means, standard deviations and correlations are given below. Complete the following table for the simple linear regression relating Height (Y) to mean foot length (X).

$$\overline{X} = 24.93 \quad \overline{Y} = 177.65 \quad s_X = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}} = 0.903 \quad s_Y = 4.810 \quad r_{XY} = 0.483$$

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

Q.30. Based on Model 1, with p = 1, defining $Q = \sum_{i=1}^{n} \varepsilon_i^2$:

p.30.a. Derive the normal equations. **Show all work**

p.30.b. Use the normal equations to show that $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} \hat{Y}_i e_i = 0$ **Show all work**

Q.31. For the simple regression model (scalar form): $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1,\ldots,n \quad \varepsilon_i \sim NID\left(0,\sigma^2\right)$

we get: $\hat{\beta}_1 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{S_{XX}} \right) Y_i, \quad \overline{Y} = \sum_{i=1}^{n} \left(\frac{1}{n} \right) Y_i, \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

p.31.a. Derive $E\left(\hat{\beta}_1\right), \quad E\left(\overline{Y}\right), \quad E\left(\hat{\beta}_0\right)$ **Show all work**

p.31.b. Derive $V\left(\hat{\beta}_1\right), \quad V\left(\overline{Y}\right), \quad \mathrm{COV}\left(\hat{\beta}_1, \overline{Y}\right), \quad V\left(\hat{\beta}_0\right), \quad \mathrm{COV}\left(\hat{\beta}_1, \hat{\beta}_0\right)$ **Show all work**

Q.32. A study in Talminadu State, India related heights (Y, cm) to left second toe length (X, cm) in a sample of n = 1020 healthy adult males. The sample means, standard deviations and correlations are given below. Complete the following table for the simple linear regression relating Height (Y) to left second toe length (X).

$$\overline{X} = 24.63 \quad \overline{Y} = 173.69 \quad s_X = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}} = 1.06 \quad s_Y = 4.57 \quad r_{XY} = 0.578$$

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