STAT 2002 - Probability and Statistics II, Summer 2024
Homework 5 - Linear Regression
100 points total.

This homework is due **Beijing Time 11:59pm Tuesday, July 23rd, 2024** on BlackBoard. No late homework accepted.

Please make sure to **SHOW ALL WORK** in order to receive full credit.

1. (20 points) We would like to compare the lifetime of four brands of pens. It is suggested that the writing surface might affect lifetime, so the following lifetimes (in minutes) are collected for each brand-surface combination:

		Writing Surface		
		1	**2**	**3**
Brand of Pen	**1**	709, 659	713, 726	660, 645
	2	668, 685	722, 740	692, 720
	3	659, 685	666, 684	678, 750
	4	698, 650	704, 666	686, 733

   (a) Fill in the following ANOVA table.

	SS	df	MS	F
Row				
Column				
Interaction				
Error				
Total				

   (b) Is there an effect on the lifetime of pen due to the brand (i.e. row effect)? Conduct the appropriate test at $\alpha = 0.05$.

   (c) Is there an effect on the lifetime of pen due to the writing surface (i.e. column effect)? Conduct the appropriate test at $\alpha = 0.05$.

   (d) Is there an interaction effect on the lifetime of pen due to both the brand and the writing surface? Conduct the appropriate test at $\alpha = 0.05$.

2. (20 points) Regression analysis were used to analyze the data from a study investigating the relationship between roadway surface temperature (°F) (x) and pavement deflection (y). Summary quantities were $n = 20$, $\sum y_i = 12.75$, $\sum y_i^2 = 8.86$, $\sum x_i = 1478$, $\sum x_i^2 = 143,215.8$, $\sum x_i y_i = 1083.67$.

   (a) Calculate the least squares estimate of the slope and intercept. Graph the regression line. Estimate σ^2.

   (b) Use the equation of the fitted line to predict what pavement deflction would be observed when the surface temperature is 85F.

   (c) What is the mean pavement deflection when the surface temperature is 90F.

(d) What change in mean pavement deflection would be expected for a 1F change in surface temperature.

3. (40 points) **Simple linear regression: NFL games.** This problem is about analyzing the NFL player performance. R instructions are provided at the end of this problem. The following table presents data on the ratings of quarterbacks for the 2008 National Football League season. It is suspected that the rating (y) is related to the average number of yards gained per pass attempt (x).

   Part I: For this dataset,

   (a) Calculate the least square estimate of the slope and intercept. What is the estimate of σ^2? Graph the regression model.

   (b) Find an estimate of the mean rating if a quarterback averages 7.5 yards per attempt.

   (c) What change in the mean rating is associated with a decrease of one yard per attempt?

   (d) To increase the mean rating by 10 points, how much increase in the average yards per attempt must be generated?

Player		Team	Yards per Attempt	Rating Points
Philip	Rivers	SD	8.39	105.5
Chad	Pennington	MIA	7.67	97.4
Kurt	Warner	ARI	7.66	96.9
Drew	Brees	NO	7.98	96.2
Peyton	Manning	IND	7.21	95
Aaron	Rodgers	GB	7.53	93.8
Matt	Schaub	HOU	8.01	92.7
Tony	Romo	DAL	7.66	91.4
Jeff	Garcia	TB	7.21	90.2
Matt	Cassel	NE	7.16	89.4
Matt	Ryan	ATL	7.93	87.7
Shaun	Hill	SF	7.10	87.5
Seneca	Wallace	SEA	6.33	87
Eli	Manning	NYG	6.76	86.4
Donovan	McNabb	PHI	6.86	86.4
Jay	Cutler	DEN	7.35	86
Trent	Edwards	BUF	7.22	85.4
Jake	Delhomme	CAR	7.94	84.7
Jason	Campbell	WAS	6.41	84.3
David	Garrard	JAC	6.77	81.7
Brett	Favre	NYJ	6.65	81
Joe	Flacco	BAL	6.94	80.3
Kerry	Collins	TEN	6.45	80.2
Ben	Roethlisberger	PIT	7.04	80.1
Kyle	Orton	CHI	6.39	79.6
JaMarcus	Russell	OAK	6.58	77.1
Tyler	Thigpen	KC	6.21	76
Gus	Freotte	MIN	7.17	73.7
Dan	Orlovsky	DET	6.34	72.6
Marc	Bulger	STL	6.18	71.4
Ryan	Fitzpatrick	CIN	5.12	70
Derek	Anderson	CLE	5.71	66.5

Part II: For this dataset,

(a) Test for significance of regression (i.e., testing whether slope is zero) using $\alpha = 0.01$. Find the p-value for this test. What conclusions can you draw?

(b) Estimate the standard errors of the slope and intercept estimators.

(c) Test $H_0 : \beta_1 = 10$ versus $H_1 : \beta_1 \neq 10$ with $\alpha = 0.01$.

Part III: Find a 95% confidence interval on each of the following:

(a) Slope β_1

(b) Intercept β_0

(c) Mean rating when the average yards per attempt is 8.0.

(d) Find a 95% confidence interval on the rating when the average yards per attempt is 8.0.

Getting the Data: Data in this problem is contained in file data113.txt. You may also use data113.csv. Once you have saved the data file in the working directory, read the data in R using the command

data = read.csv("data113.csv",header=TRUE)

We can investigate the association of `Rating` to `Yds` using linear regression. Define first

```
y = data$Rating
x = data$Yds
```

The function to fit a linear regression model in R is `lm`. We perform a linear regression with R as follows

```
model = lm(y~x)
summary(model)
```

You can calculate by hand, or use the fitted results to help you answer above questions. If you use R, make sure to include the R output in your homework and state clearly how do you use the R output to help solve the above questions.

4. (20 points) **Multiple linear regression.** The electric power consumed each month by a chemical plant (y) is related to the average ambient temperature x_1, the number of days in the month x_2, the average product purity x_3, and the tons of product produced x_4. The past year's historical data are available and are presented in the following table.

y	x_1	x_2	x_3	x_4
240	25	24	91	100
236	31	21	90	95
270	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

Use R to find the multiple linear regression model. For example, you may create the variable x_1 in R using the following command

```
x1 = c(25,31,45,60,65,72,80,84,75,60,50,38)
```

You can fit a multiple linear regression model similar to simple linear regression using the following command

```
model = lm(y~x1+x2+x3+x4)
summary(model)
```

Based on the results or R, answer the following questions:

(a) Fit a multiple linear regression model to these data.

(b) Estimate σ^2.

(c) Compute the standard errors of the regression coefficients. Are all of the model parameters estimated with the same precision? Why or why not?

(d) Predict the power consumption for a month in which $x_1 = 75$ F, $x_2 = 24$ days, $x_3 = 90\%$, and $x_4 = 98$ tons.