

Homework Problems 1

Applied Regression Analysis
Chicago Booth

1 Normal Probability Calculations (Section 1)

$X \sim N(5, 10)$ (Read X distributed Normal with mean 5 and var 10) Compute:

(i) $\text{Prob}(X > 5)$

(ii) $\text{Prob}(X > 5 + 2 \times \sqrt{10})$

(iii) $\text{Prob}(X < 4)$

(iv) $\text{Prob}(6 < X < 15)$

(v) $\text{Prob}(-2 \leq X \leq 6) = \text{Prob}(a \leq Z \leq b)$, where Z is the standard normal random variable. What are a and b ?

3 Histograms (Section 1)

For each histogram below, provide a guess for the average and sample variance.

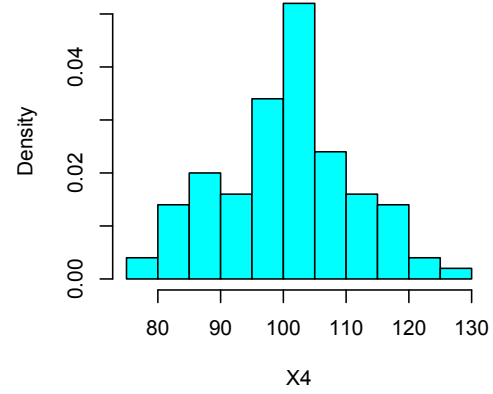
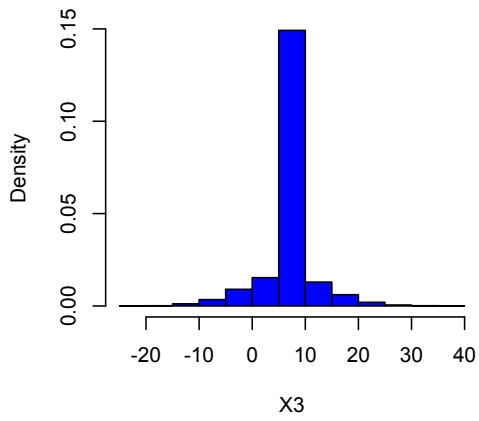
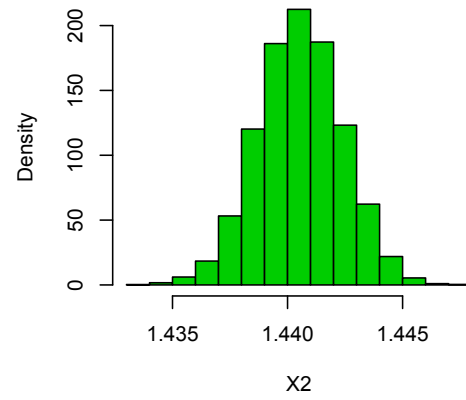
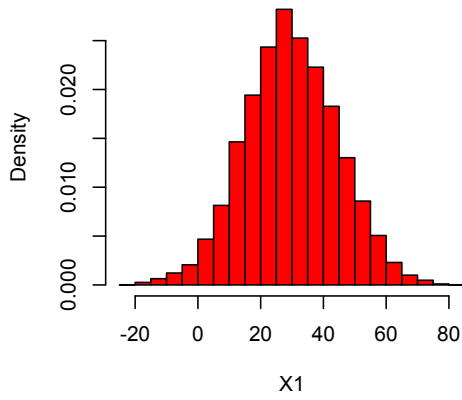


Figure 1: Histograms

4 Match the Plots (Section 1)

Below (Figure 1) are 4 different scatter plots of an outcome variable y versus predictor x followed by 4 four regression output summaries labeled A, B, C and D. Match the outputs with the plots.

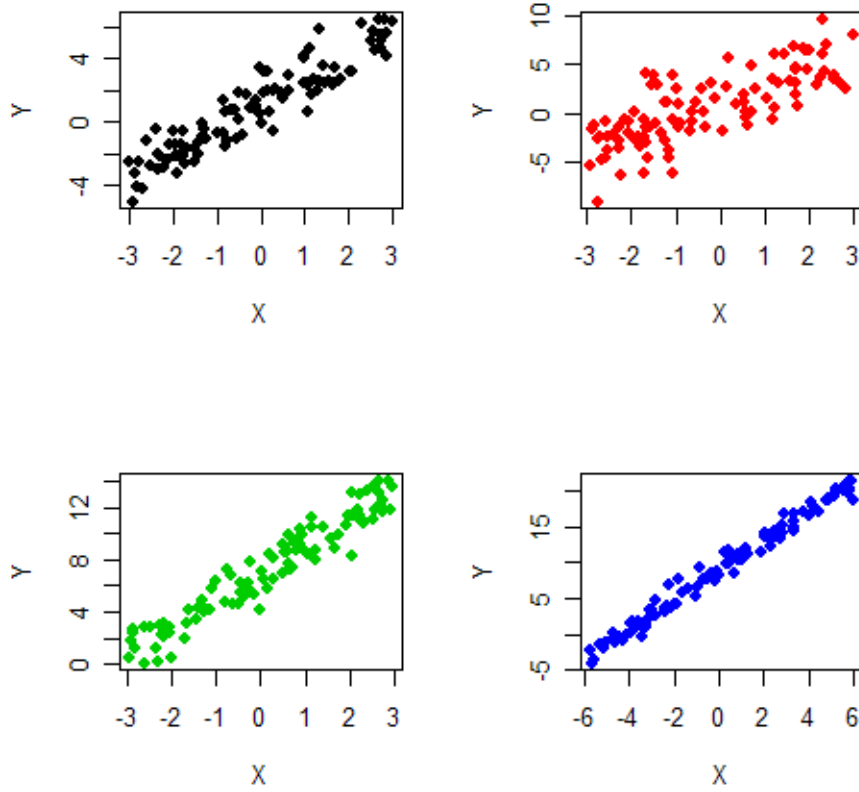


Figure 2: Scatter Plots

Regression A:

Coefficients:

	Estimate	Std. Error
(Intercept)	7.03747	0.12302
(Slope)	2.18658	0.07801

Residual standard error: 1.226

R-Squared: 0.8891

Regression B:

Coefficients:

	Estimate	Std. Error
(Intercept)	1.1491	0.1013
(Slope)	1.4896	0.0583

Residual standard error: 1.012

R-Squared: 0.8695

Regression C:

Coefficients:

	Estimate	Std. Error
(Intercept)	1.2486	0.2053
(Slope)	1.5659	0.1119

Residual standard error: 2.052

R-Squared: 0.6666

Regression D:

Coefficients:

	Estimate	Std. Error
(Intercept)	9.0225	0.0904
(Slope)	2.0718	0.0270

Residual standard error: 0.902

R-Squared: 0.9835

5 More on SLR Model (Section 1)

Suppose we are modeling house price as depending on house size. Price is measured in thousands of dollars and size is measured in thousands of square feet.

Suppose our model is:

$$P = 20 + 50s + \epsilon, \quad \epsilon \sim N(0, 15^2).$$

- (a) Given you know that a house has size $s = 1.6$, give a 95% predictive interval for the price of the house.

- (b) Given you know that a house has size $s = 2.2$, give a 95% predictive interval for the price.

- (c) In our model the slope is 50. What are the units of this number?

- (d) What are the units of the intercept 20?

- (e) What are the units of the error standard deviation 15?

- (f) Suppose we change the units of price to dollars and size to square feet
What would the values and units of the intercept, slope, and error standard deviation?

- (g) If we plug $s = 1.6$ into our model equation, P is a constant plus the normal random variables ϵ . Given $s = 1.6$, what is the distribution of P ?

6 The Shock Absorber Data (Section 1)

The data comes from a company which supplies a major automobile manufacturer with shock absorbers. An important characteristic is the “force transferred through the shock absorber when the shank is forced out of the cylinder”. If you don’t know what that really means, don’t worry, neither do I.

What we do need to understand is that the manufacturer only considers the shock to be an acceptable part if the force measurement is between 485 and 585.

The shock manufacturer and the auto manufacturer are arguing over the following issue. Before the shock is finally shipped, it is filled with gas. After it is filled with gas, it becomes very difficult to measure the force characteristic we are interested in. The shock manufacturers would like to make the measurement before the shock is filled with gas. The auto maker is concerned that there may be a difference in the force before and after the shock is filled with gas and so would like to make the measurement after it is filled.

The shock maker claims that there is little difference between the before and after measurement so that the before measurement can be used.

To investigate this we have the before (column 1, reboundb) and the after (column 2, rebounda) measurements on 35 shocks. The data for this problem is in **shock.csv** available in the class website.

- (a) Plot the before measurement vs. the after measurement. Does this look like the kind of data we can use the simple linear regression model to think about? Why does it make sense to choose the after measurement as “Y” and the before measurement as “X”?
- (b) What are 95% confidence intervals for both the slope and the intercept?
- (c) Test the null hypothesis $H_0 : \beta_0 = 0$.
- (d) From the shock maker’s point of view, what hypotheses would be of interest to test for the slope and intercept. That is, what would the shock maker like the true intercept and slope to be?
 - Test whether the intercept is equal to the value proposed by the shock maker.
 - Test whether the slope is equal to the value proposed by the shock maker.
- (f) Suppose the before measurement is 550.
 - What is the plug-in predictive interval given $x\text{-before}=550$.
 - What does this interval suggest about the acceptability of the shock absorber?

7 SLR (Section 1)

Let Y and X denote the labor force participation rate of women in 1972 and 1968, respectively, in each of 19 cities in the US. The regression output for this data set is shown in the Table below:

Variable	Coefficient	s.e.	t-test	p-value
Intercept	0.203311	0.0976	2.08	0.0526
X	0.656040	0.1961	3.35	0.0036
$n = 19$	$R^2 = 0.397$		$s = 0.0566$	$d.f. = 17$

It was also found that $SSR = 0.0358$ and $SSE = 0.0544$. Suppose that the model satisfies the usual regression assumptions.

- (a) Compute the sample variance of Y and $Corr(Y, X)$.
- (b) Suppose that the participation rate of women in 1968 in a given city is 45%. What is the estimated participation rate of women in 1972 for the same city?
- (c) Construct the 95% confidence interval for the slope of the true regression line β_1 .
- (d) Test the hypothesis $\beta_1 = 1$ (with 95% probability).
- (e) If Y and X were reversed in the above regression, what would you expect R^2 to be?

9 Market Model Example (Section 1)

The *Market Model*, a well-known model in finance assumes that the rate of return R^g on a generic stock is linearly related to the rate of return (R^m) on the overall stock market as:

$$R_i^g = \alpha + \beta R_i^m + \epsilon_i$$

where the error term ϵ follow the assumptions of the Simple Linear Regression Model. For practical purposes, R^m is taken to be the rate of return on some major stock-market index (such as the S&P500). The slope coefficient β , called the *beta* of the stock, measures the sensitivity of the stock's rate of return to changes in the level of the overall market.

The file **mktmodel.csv** contains 60 monthly returns (from 1992 to 1996) of the S&P500 index and 30 companies.

- (a) Choose any three companies from the list and run the individual market-model regressions. Then,
1. Compare the slope estimates from the three regressions and interpret.
 2. Compare the intercept estimates from the three regressions. Does this comparison tell you anything useful?
 3. What is the 95% confidence interval for β in the regressions?
 4. For one of the selected companies, test the hypothesis (with 99% probability) that $\beta = 1$? What is your conclusion? How would you interpret this result in simple financial terms?

10 Simulation from the SLRM (Section 1)

To drive home the concepts of the regression probability model, run a simulation exercise with the following steps:

(a) Generate 70 samples of $X \sim N(0, 2)$

(b) Simulate 70 samples of Y from the simple linear regression model.

$$Y_i = 3.0 + 1.0X_i + \epsilon_i \quad \epsilon_i \sim N(0, 2^2)$$

(c) Show the scatter plot of Y versus X along with the true regression line.

(d) Change the distribution of ϵ to $N(0, 4^2)$.

(e) Show the new scatter plot of Y versus X along with the true regression line. Is the true regression line different than the one in (c)?

(f) Comment on the difference between the scatter plots of (c) and (e).

(g) Split the sample from item (b) in 3 subsets of size 10, 20 and 40. For each subset, run the regression of Y on X . Show the 3 scatter plots of each of the subsets along with the fitted regression line. Are the lines the same? Which line is closest to the true regression line? Briefly describe the output of the three regressions and your understanding of the results.

(h) Suppose I told you $X = -1$. From the true model in (b), what is your best guess for Y ? What is the 68%, 75% and 95% predictive intervals for Y ?

11 Sampling Distribution (Section 1)

In class we learned that given the simple linear model

$$Y = \beta_0 + \beta_1 X + \epsilon, \epsilon \sim N(0, \sigma^2),$$

the distribution for the estimates b_0 and b_1 are given by

$$b_0 \sim N(\beta_0, s_{b_0}^2) \text{ and } b_1 \sim N(\beta_1, s_{b_1}^2).$$

As a skeptical student you wonder if this is actually the case. To see if Professor Carvalho was indeed telling you the truth you decide to run a simulation.

In particular, you remember that $b_0 \sim N(\beta_0, s_{b_0}^2)$ is the same as saying that $t_{b_0} = \frac{b_0 - \beta_0}{s_{b_0}} \sim N(0, 1)$, which means that to check that b_0 is distributed as $b_0 \sim N(\beta_0, s_{b_0}^2)$ you can check that $t_{b_0} \sim N(0, 1)$. An analogous statement holds for t_{b_1} . Thus you decide to run a simulation that produces samples of t_{b_0} and t_{b_1} and check if these appear to be $N(0, 1)$ to see if the Professor was telling the truth.

Here is an outline of what you need to do:

1. Pick some “true” values for the parameters β_0 , β_1 , and σ . Pick some “true” values for $X_i, i = 1, \dots, 50$. It does not matter what they are.
2. Generate some “data” $Y_i, i = 1, \dots, 50$ using the model

$$Y = \beta_0 + \beta_1 X + \epsilon, \epsilon \sim N(0, \sigma^2),$$

3. Using this data, perform a linear regression to get the values of b_0 and b_1 .
4. Calculate the value of s , s_{b_0} , and s_{b_1} . Look up the formulas in the notes. You will have to use the residuals provided by the regression output to calculate s .
5. Calculate the value of t_{b_0} and t_{b_1} and save them...
6. Repeat steps (b) - (e) many times. Check out the histograms of the values obtained for t_{b_0} and t_{b_1} .

If you are daring you will repeat this test when there are only $n = 5$ data points.

12 The Multiple Linear Regression Model (Section 2)

Suppose we are modeling house price as depending on house size, the number of bedrooms in the house and the number of bathrooms in the house. Price is measured in thousands of dollars and size is measured in thousands of square feet.

Suppose our model is:

$$P = 20 + 50 \text{ size} + 10 \text{ nbed} + 15 \text{ nbath} + \epsilon, \quad \epsilon \sim N(0, 10^2).$$

(a) Suppose you know that a house has size =1.6, nbed = 3, and nbath =2.

What is the distribution of its price given the values for size, nbed, and nbath.

(hint: it is normal with mean = ?? and variance = ??)

(b) Given the values for the explanatory variables from part (a), give the 95% predictive interval for the price of the house.

(c) Suppose you know that a house has size =2.6, nbed = 4, and nbath =3. Give the 95% predictive interval for the price of the house.

(d) In our model the slope for the variable nbath is 15. What are the units of this number?

(e) What are the units of the intercept 20? What are the units of the the error standard deviation 10?

13 Profits... (Section 2)

For this problem us the data is the file **Profits.csv**.

There are 18 observations.

Each observation corresponds to a project developed by a firm.

y = Profit: profit on the project in thousands of dollars.

x_1 = RD: expenditure on research and development for the project in thousands of dollars.

x_2 = Risk: a measure of risk assigned to the project at the outset.

We want to see how profit on a project relates to research and development expenditure and “risk”.

- (a) Plot profit vs. each of the two x variables. That is, do two plots y vs. x_1 and y vs x_2 . You can't really understand the full three-dimensional relationship from these two plots, but it is still a good idea to look at them. Does it seem like the y is related to the x 's?
- (b) Suppose a project has risk=7 and research and development = 76. Give the 95% plug-in predictive interval for the profit on the project. Compare that to the correct, predictive interval (using the predict function in R).
- (c) Suppose all you knew was risk=7. Run the simple linear regression of profit on risk and get the 68% plug-in predictive interval for profit.
- (d) How does the size of your interval in (c) compare with the size of your interval in (b)? What does this tell us about our variables?

14 Zagat... (Section 2)

The data for this question is in the file `zagat.csv` . The data is from the Zagat restaurant guide. There are 114 observations and each observation corresponds to a restaurant.

There are 4 variables:

price: the price of a typical meal

food: the zagat rating for the quality of food.

service: the zagat rating for the quality of service.

decor: the zagat rating for the quality of the decor.

We want to see how the price of a meal relates the quality characteristics of the restaurant experience as measured by the variables food, service, and decor.

- (a) Plot price vs. each of the three x 's. Does it seem like our y (price) is related to the x 's (food, service, and decor) ?
- (b) Suppose a restaurant has food = 18, service=14, and decor=16. Run the regression of price on food, decor, and service and give the 95% predictive interval for the price of a meal.
- (c) What is the interpretation of the coefficient estimate for the explanatory variable food in the multiple regression from part (b) ?
- (d) Suppose you were to regress price on the one variable food in a simple linear regression? What would be the interpretation of the slope? Plot food vs. service. Is there a relationship? Does it make sense? What is your prediction for how the estimated coefficient for the variable food in the regression of price on food will compare to the estimated coefficient for food in the regression of price on food, service, and decor? Run the simple linear regression of price on food and see if you are right! Why are the coefficients different in the two regressions?
- (e) Suppose I asked you to use the multiple regression results to predict the price of a meal at a restaurant with food = 20, service = 3, and decor =17. How would you feel about it?

15 More on Multiple Regression Model (Section 2)

The following table shows the regression output of a multiple regression model relating the beginning salaries in dollars of employees in a given company to the following predictor variables:

- Sex: An indicator variable (1=man and 0=woman);
- Education: Years of schooling at the time of hire;
- Experience: Number of months of previous work experience;
- Months: Number of months with the company.

ANOVA Table

Source	Sum of Squares	d.f.	Mean Square	F-test
Regression	23665352	4	5916338	22.98
Residuals	22657938	88	257477	

Coefficients Table

Variable	Coefficient	s.e.	t-test	p-value
Constant	3526.4	327.7	10.76	0.000
Sex	722.5	117.8	6.13	0.000
Education	90.02	24.69	3.65	0.000
Experience	1.2690	0.5877	2.16	0.034
Months	23.406	5.201	4.50	0.000
$n = 93$	$R^2 = 0.515$	$R_a^2 = 0.489$	$s = 507.4$	

1. Conduct a F -test for the overall fit of the regression
2. Is there a *positive* linear relationship between Salary and Experience, after accounting for the effect of the variables Sex, Education and Months?
3. What salary would you forecast for a man with 12 years of education, 10 months of experience and 15 months with the company?

Now consider a reduced model in which Salary is regressed on Education only. The ANOVA table obtained when fitting this model is shown below. Conduct a single test to compare the full and reduced models at a $\alpha = 0.05$ level. What conclusion can be drawn from the result of the test?

ANOVA Table

Source	Sum of Squares	d.f.	Mean Square	F-test
Regression	7862535	1	7862535	18.60
Residuals	38460756	91	422646	

16 Baseball (Section 3)

Using our baseball data (**RunsPerGame.csv**), regress R/G on a binary variable for league membership (League = 0 if National and League = 1 if American) and OBP .

$$R/G = \beta_0 + \beta_1 League + \beta_2 OBP + \epsilon$$

1. Based on the model assumptions, what is the expected value of R/G given OBP for teams in the AL? How about the NL?
2. Interpret β_0 , β_1 and β_2 .
3. After running the regression and obtaining the results, can you conclude with 95% probability that the marginal effect of OBP on R/G (after taking into account the League effect) is positive?
4. Test the hypothesis that $\beta_1 = 0$ (with 99% probability). What do you conclude?

17 Corn Yields and Fertilizers (Section 3)

Three types of fertilizers are to be tested to see which one yields more corn crop. Forty similar plots of land were available for testing purposes. The 40 plots were divided at random into 4 groups, ten for each group. Fertilizer 1 was applied to the plots in group 1, and Fertilizers 2 and 3 were applied to groups 2 and 3. The corn plants in group 4 were not given any fertilizer and will serve as the control group. The table below gives the corn yields for each of the forty plots:

Fertilizer 1	Fertilizer 2	Fertilizer 3	Control Group
31	27	36	33
34	27	37	27
34	25	37	35
34	34	34	25
43	21	37	29
35	36	28	20
38	34	33	25
36	30	29	40
36	32	36	35
45	33	42	29

- (a) Create 3 dummy variables X_1 , X_2 and X_3 one for each fertilizer groups.
- (b) Test the hypothesis that none of the three fertilizers has an effect on corn crops. Define the model, specify the hypothesis to be tested, the test used and your conclusion at the 5% significance level.
- (c) Test the hypothesis that the use of fertilizers (any one of the 3) have a positive effect on corn crops. Define the model, specify the hypothesis to be tested, the test used and your conclusion at the 5% significance level.
- (d) Based on a multiple regression decide what fertilizer has the greatest effect on corn yield.

18 Housing Price Structure (Section 3)

The file **MidCity.csv**, available on the class website, contains data on 128 recent sales of houses in a town. For each sale, the file shows the neighborhood in which the house is located, the number of offers made on the house, the square footage, whether the house is made out of brick, the number of bathrooms, the number of bedrooms, and the selling price. Neighborhoods 1 and 2 are more traditional whereas 3 is a more modern, newer and more prestigious part of town. Use regression models to estimate the pricing structure of houses in this town. Consider, in particular, the following questions and be specific in your answers:

1. Is there a premium for brick houses everything else being equal?
2. Is there a premium for houses in neighborhood 3?
3. Is there an extra premium for brick houses in neighborhood 3?
4. For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single “older” neighborhood?