Practice Problems

1. A survey asked 683 students how many hours they spent doing homework in a week. On average, the 292 surveyed men responded 8.4 hours. The standard deviation of their responses was 9.5. On average, the 391 surveyed women responded 12.8 hours, with standard deviation of 11.6. We are interested in testing whether men spend less hours on homework than women. (Note: round all numbers to the nearest thousandth).

   a. Formulate the null and alternative hypotheses.

   H_0: $\mu_{women} - \mu_{men} = 0$
   H_a: $\mu_{women} - \mu_{men} > 0$

   b. Calculate the standard error for this hypothesis test

   For a large sample:

   $SE = \sqrt{(S_1^2/n_1 + S_2^2/n_2)} = \sqrt{(9.5^2/292 + 11.6^2/391)} = \sqrt{(.309 + .344)} = .808$

   c. With a confidence level of 99%, what would you conclude?

   $$z = \frac{\mu_{women} - \mu_{men} - 0}{SE} = \frac{12.8 - 8.4}{.808} = 5.446$$

   $P(5.446) = 2.576 \times 10^{-8}$ (we DO NOT multiply by two, as this is a one-sided test. At a 99% confidence level ($\alpha = .01$), we can reject the null hypothesis: it appears that men and women do spend different amounts of time on homework.

   d. Construct a 99% confidence interval for the difference in the two means. Interpret.

   Z(99% CI) = 2.576

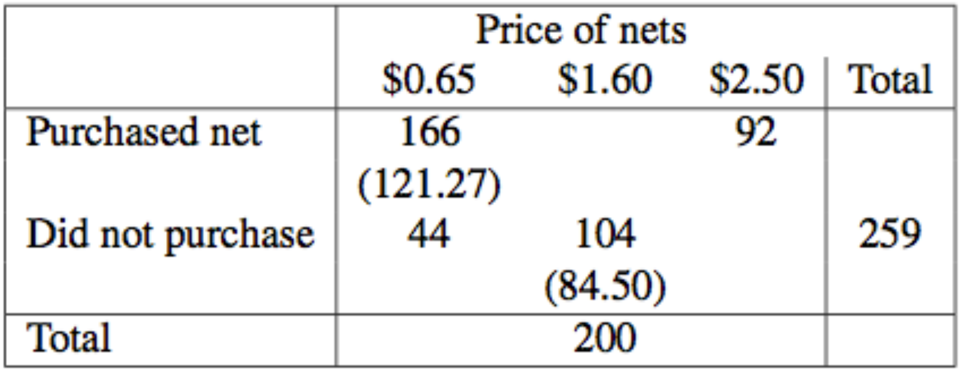
   $\mu_{women} - \mu_{men} \pm z(SE) = 4.4 \pm 2.576(.808) = 4.4 \pm 2.081$ CI = [2.319, 6.481]

   With repeated random sampling of 683 students, 99% of the differences in means between men and women will fall between 2.319 and 6.481.

   e. What is the relationship between the p-value you observed in part c, and the confidence interval calculated in part d?

The p-value proved that the difference between men's and women's time spent on homework was not 0 at a 99% confidence level. The confidence interval proved this—the 99% confidence interval for the difference between men's and women's time spent on homework does not include 0.

2. A group of researchers are conducting experiments to see how the price of mosquito nets sold at hospitals affects the number of patients who choose to purchase them. Unfortunately, the researchers have had computer trouble, and they have only been able to retrieve the information presented in the following table. It contains partial information for each cell, including some observed counts, some expected frequencies (in parentheses), and some column and row totals.



	Price of nets			
	$0.65	$1.60	$2.50	Total
Purchased net	166 (121.27)		92	
Did not purchase	44	104 (84.50)		259
Total		200		

a. Use the information listed to complete the table. Be sure to calculate both the observed and expected frequencies for each cell.

	Price of nets			
	$0.65	$1.60	$2.50	Total
Purchased Net	**166** **(121.27)**	200-104=**96** (354x200)/613=**115.498**	**92** (354x203)/613=**117.230**	166+96+92=**354**
Did not purchase	**44** (259x210)/613=**88.727**	**104** **(84.50)**	259-44-104=**111** (259x203)/613=**85.770**	**259**
Total	166+44=**210**	96+104=**200**	92+111=**203**	259+354=**613**

(Row Total x Column Total)/Grand Total = Expected Total

b. Calculate the standardized residual for the upper-left cell of the table (i.e., Purchased nets when the price was $0.65). Interpret this statistic.

Z= f_0-f_e/√[f_e(1-row prop.)(1-column prop.)]

Row prop. = 354/613 = .577
Column prop. = 210/613 = .343
= 166-121.27/√[121.27(1-.577)(1-.343)] = 44.73/√121.27(.423)(.657) = 44.73/5.805 = 7.705

Because the standardized residual is greater than 0 and relatively large, the frequency in this cell is significantly higher than it would be if there was no relationship between the two variables.

c. Calculate the χ^2 statistic for this table. Specify and conduct a hypothesis test using this number. Interpret your results. What do these results tell us about the relationship between price and the purchasing of nets?

H_0: there is no association between the variables
H_a: there is an association between the variables

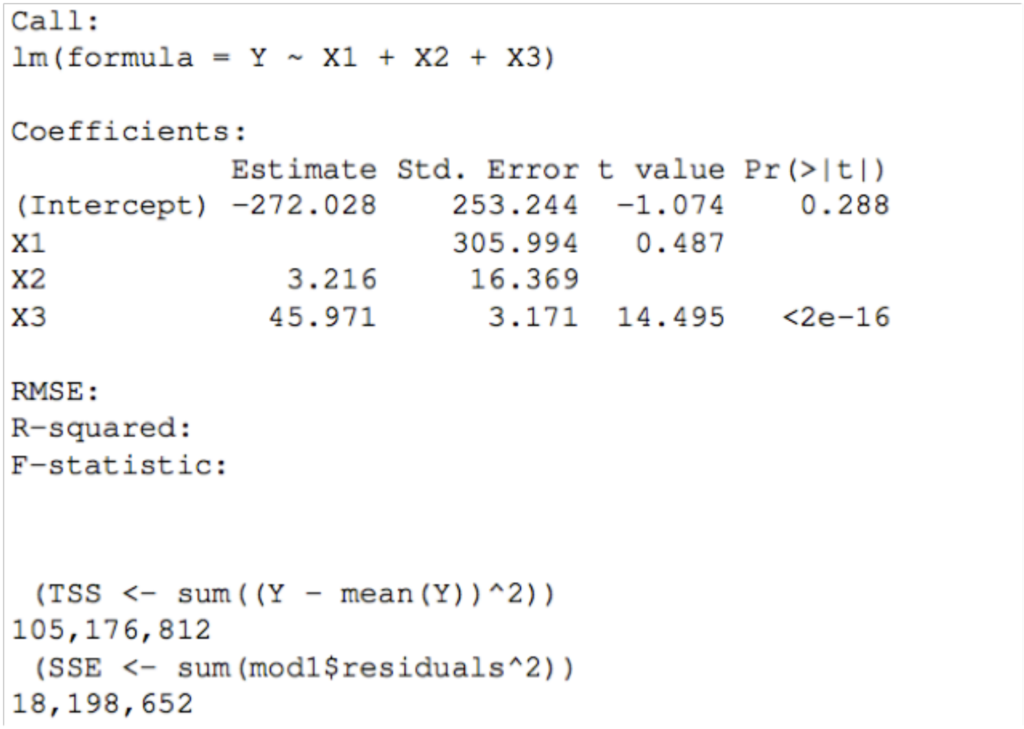
$\chi^2 = \Sigma(f_o-f_e)^2/f_e = 16.498 + 3.292 + 5.430 + 22.547 + 4.500 + 7.422 = 59.689$

df = (rows-1)(columns-1) = (1)(2) = 2

$p = 1.093\text{x}10^{-13} \approx 0$

This means that we reject the null hypothesis: there is an association between these two variables.

3. For the regression output below, n = 50.



```
Call:
lm(formula = Y ~ X1 + X2 + X3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -272.028    253.244  -1.074    0.288
X1                      305.994   0.487
X2             3.216     16.369
X3            45.971      3.171  14.495   <2e-16

RMSE:
R-squared:
F-statistic:



 (TSS <- sum((Y - mean(Y))^2))
105,176,812
 (SSE <- sum(mod1$residuals^2))
18,198,652
```

Find:

a. The p-value for the effect of X1 on Y
Df = n – (k + 1) = 46

P(.487) = .314(2) = .618

b. The test statistic for the null hypothesis that X2 has no effect on Y. Find the p-value and interpret.

t = β/SE = 3.216/16.369 = .196
df = 50 – (3 + 1) = 46
P(.196) = .423(2) = .846. At any normal level of significance, we cannot reject the null hypothesis that X2 has no effect on Y—it is possible that X2 has no effect on Y.

c. The RMSE

RMSE = √SSE/[n – (k + 1)] = √[18,198,652/(50 – 4)] = 628.986

d. The R^2 value

$$\mathbf{R^2 = \frac{\underline{TSS - SSE}}{TSS}}$$

$$\mathbf{= \frac{\underline{105{,}176{,}812 - 18{,}198{,}652}}{105{,}176{,}812} = .827}$$

e. A 95% confidence interval for the estimated effect of X3 on Y

df = n – (k+1) =50 – 4 = 46
β ± t_{46}(SE)
45.971 ± 2.013(3.171) = 45.971 ± 6.383 = [39.588, 52.354]

f. The X1 estimate

t = β/SE

.487 = β/305.994

β = 149.019

g. The F Statistic. Calculate the p-value and interpret.

$$\mathbf{F = \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]} = \frac{.827/3}{(1-.827)/[50 - (3 + 1)]}}$$

= .276/.00376 = 73.387

p(73.386 on 3 and 46 df) = 1.466 x 10^{-17} ≈ 0. So we can reject the null hypothesis—the multiple correlation ≠ 0.

h. Write the multiple regression equation.

Y = -272.028 + 149.019X1 + 3.216X2 +45.971X3

4. In 2007, 11 South American countries had an average GDP per capita of $10,752 with a standard deviation of $6,500. 14 Central American countries had an average GDP per capital of $12,936 with a standard deviation of $3,500. Are income levels different between these two regions?

H_0: $\mu_2 - \mu_1 = 0$
H_a: $\mu_2 - \mu_1 \neq 0$

σ_{hat} = √[(11-1)(6500^2) + (14-1)(3500^2)]/(11 + 14 – 2)
= 5029.262

SE = 5029.262√(1/11 + 1/14) = 2026.347

$$\mathbf{t = \frac{12{,}936 - 10{,}752 - 0}{2026.347} = 1.078}$$

df = 11 + 14 – 2 = 23

p(1.078, df = 23) = 0.146(2) = .292

At a reasonable level of confidence, we cannot reject the null hypothesis: GDP per capita is not necessarily different between Central and South America

5. The National Health Interview Survey estimated that current cigarette smokers were 41.9% of American adults in 1965 and 21.5% in 2003. In 1965, the sample size was 4,238 and in 2003 the sample size was 6,947.

   a. Does the percentage of smokers in 2003 differ from the percentage of smokers in 1965 at a significance level of 0.01?

      $\mathbf{H_0: \pi_2 - \pi_1 = 0}$
      $\mathbf{H_a: \pi_2 - \pi_1 \neq 0}$

      $$\boldsymbol{\pi_{pooled} = \frac{\underline{.419(4238) + .215(6947)}}{4238 + 6947} = .292}$$

      $\mathbf{SE_{pooled} = \surd .292(1-.292)/4238 + .292(1-.292)/6947 = .0089}$

      $$\mathbf{Z = \frac{\underline{.419-.215-0}}{.0089} = 22.921}$$

      p(22.921) *2 ≈ 0. Therefore, we can reject the null hypothesis: the percentage of smokers in 2003 does differ from the percentage of smokers in 1965.

   b. Construct a 99% confidence interval around the difference in the two proportions.

      .419 - .215 = .204

      $\mathbf{SE = \surd .419(1-.419)/4238 + .215(1-.215)/6947 = .009}$

      $\mathbf{.204 \pm 2.576(.009) = .204 \pm .023 = [.181, .227]}$

      Because this confidence interval does not include 0, it supports the findings in our hypothesis test.

6. Researchers interested in determining if there is a relationship between death anxiety and religiosity conducted the following study. Subjects completed a death anxiety scale (high score=high anxiety) and also completed a checklist designed to measure an individual's degree of religiosity (high score=greater religiosity. For simplicity's sake, assume that

religiosity is the explanatory (independent) variable and death anxiety is the response (dependent) variable. Use the data provided to answer the questions below.

X (religiosity)	Y (death anxiety)	(X_i-X_{bar})	(Y_i-Y_{bar})	$(X_i-X_{bar})*(Y_i-Y_{bar})$	$(X_i-X_{bar})^2$	$(Y_i-Y_{bar})^2$
14	24	1.625	5.125	8.328	2.641	26.266
5	31	-7.375	12.125	-89.422	54.391	147.016
10	19	-2.375	.125	-.297	5.641	0.016
18	6	5.625	-12.875	-72.422	31.641	165.766
19	8	6.625	-10.875	-72.047	43.891	118.266
4	38	-8.375	19.125	-160.172	70.141	365.766
14	14	1.625	-4.875	-7.922	2.641	23.766
15	11	2.625	-7.875	-20.672	6.891	62.016
X_{bar}=12.375	Y_{bar}=18.875			**Total** -414.626	**Total** 217.878	**Total** 908.878

a. Calculate *β*.

β = -414.626/217.878 = -1.903

b. Calculate *α*.

α = 18.875 – (-1.903)(12.375) = 42.425

c. Write the regression equation and substantively interpret each part of the equation.

y = 42.425 – 1.903x

y is death anxiety; x is religiosity. 42.425 is the y-intercept: if a person is not at all religious, her death anxiety score will be 42.425. -1.903 is the slope—for every 1 unit increase in religiosity, death anxiety will decrease by 1.903 units.

d. Calculate conditional and unconditional standard deviation

Unconditional: use y_{bar}. √908.878/(8-1) = 11.395

Conditional: calculate y_{hat}:

X (religiosity)	**Y (death anxiety)**	**Y_{hat} (y_{hat} = 42.425 – 1.903x)**	**$(Y-Y_{hat})^2$**
14	**24**	**15.783**	**67.519**
5	**31**	**32.91**	**3.648**
10	**19**	**23.395**	**19.316**
18	**6**	**8.171**	**4.713**
19	**8**	**6.268**	**3.000**
4	**38**	**34.813**	**10.157**
14	**14**	**15.783**	**3.179**

15	**11**	**13.88**	**8.294**
			Total: 119.826

$\sqrt{}119.826/6 = 4.469$

e. Find TSS

TSS = 908.878

f. Find SSE

SSE = 119.826

g. Calculate *r*

$r = (s_x/s_y)\beta$

$s_x = \sqrt{}217.878/7 = 5.579$

$s_y = \sqrt{}908.878/7 = 11.395$

$r = (5.579/11.395)(-1.903) = -.932$

h. Calculate *r^2*

$$r^2 = \frac{TSS\text{-}SSE}{TSS} = \frac{908.878 - 119.827}{908.878} = .868$$

i. Does a correlation exist between religiosity and death anxiety? Conduct a hypothesis test to find out.

H_0: $\beta = 0$
H_a: $\beta \neq 0$

$t = \beta/SE$

$SE = 4.469/\sqrt{}217.878 = .303$

$t = -1.903/.303 = -6.281$

$p(-6.281, df = n-2 = 6) = .000379(2) = .000758$

So we reject the null hypothesis: a correlation does exist between religiosity and death anxiety.

7. Below is a logit function to model admission into UCLA graduate programs:

$$\text{Logit}[P(\text{Y} = 1)] = -4.949 + 0.0026*\text{GRE} + 0.757*\text{GPA}$$

Find the estimated probability of being admitted to graduate school for:

a. An applicant with a GRE score of 750 and a GPA of 4.0

$$\boldsymbol{\frac{\exp(-4.949 + .0026(750) +.757(4.0))}{1 + \exp(-4.949 + .0026(750) +.757(4.0))}}$$

$= e^{.029}/(1 + e^{.029}) = .507$

b. An applicant with a GRE score of 450 and a GPA of 2.0

$$\boldsymbol{\frac{\exp(-4.949 + .0026(450) +.757(2.0))}{1 + \exp(-4.949 + .0026(450) +.757(2.0))}}$$

$= e^{-2.265}/(1 + e^{-2.265}) = .094$

c. If the standard error for the GPA coefficient is .322, conduct a hypothesis test for the GPA coefficient using a significance level of 0.05. Interpret the results.

H_0: GPA has no effect on grad school entrance ($\beta_{GPA} = 0$)
H_a: GPA has an effect on grad school entrance ($\beta_{GPA} \neq 0$)

$z = .757/.322 = 2.351$

$p(2.351) = .00936(2) = .0187$

So, we can reject the null hypothesis at a significance level of 0.05: GPA has an effect on grad school entrance.

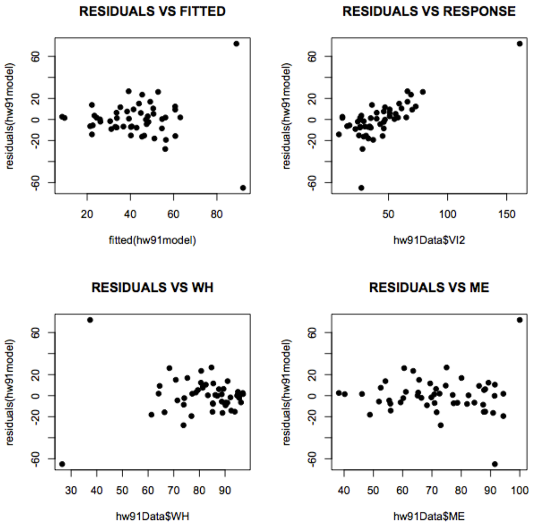
8. In the plots below, the residuals of a regression model are plotted against several different quantities. The regression model predicts violent crime rates for states based on several covariates. WH is one of these covariates, as is ME (percent of people living in metropolitan areas).

   "Residuals vs Fitted" plots the residuals for the regression against the fitted (expected) values for each point. "Residuals vs Response" plots the residuals against the response (outcome) variable). "Residuals vs WH" and "Residuals vs ME" plot the residuals against each of the covariates.

   a) What is indicated by the positive relationship between the residuals and the response variable?
   As the response variable increases, the residual value increases as well—this suggests that the errors are not independent of the response variable.

   b) Is there evidence of any outliers? If so, do you think those outliers seem to be high leverage? Explain your answer.
   Yes. In each graph two outliers are easily visible in the corners. The outliers appear to be high leverage because the residuals seem to spread away from each other in the direction of the outliers—as the line of best fit attempted to capture the outliers, it got further away from the other data points, causing the magnitude of the residuals to increase.

   c) Do these residuals indicate any other potential violations of standard regression assumptions?
   They seem to violate the assumption of homoscedasticity—the variance of the errors is not constant (this is visible in the residuals vs. fitted graph, for example). Based on this and the suggestion that the errors are not independent of the response variable, we can guess that the assumption of normality in the distribution of the errors is being violated.



9. A Pew Research Center poll (May 14, 2003) of 1201 adults asked, “All in all, do you think affirmative action programs designed to increase the number of black and minority students on college campuses are a good thing or a bad thing?” 60% said good, 30% said bad, and 10% didn’t know. Does this data prove that a majority of people are in favor of affirmative action? Conduct a hypothesis test using a significance level of 0.05.

H_0: $\pi_{favor} = .5$
H_a: $\pi_{favor} > .5$

SE = $\sqrt{.5(1-.5)/1201}$ = .0144

$z = \frac{.6 - .5}{.0144} = 6.944$

p (6.944) = $1.91 \times 10^{-12} \approx 0$
So, we can reject the null hypothesis: the proportion of people who favor affirmative action is greater than 50%

10. For recent data in Florida on y = selling price of a home (in dollars), x_1 = size of home (in square feet), x_2 = lot size (in square feet), the prediction equation is:

$$y = -10,536 + 53.8x_1 + 2.84x_2$$

a. Predict the selling price for a home of 1240 square feet on a lot of 18,000 square feet.

y = -10,536 + 53.8(1240) + 2.84(18,000) = $107,296

b. For fixed lot size, how much is the house’s selling price predicted to increase for each square-foot increase in home size? Why?

$53.80, the coefficient for the home size variable. If lot size is fixed, we can take it out of the equation. A 1-square foot increase is $53.80(1).

11. The table below is a regression output for the selling price of 32 grandfather clocks sold at auction. “Age” represents age of the clock (in years), “Bidders” represents the number of individuals participating in bidding, and “Price” represents selling price (in pounds sterling).

Predictor	**Coefficient**	**SE for Coefficient**
Constant	322.8	293.3
Age	0.873	2.020
Bidders	-93.41	29.71
Age*Bidders	1.2979	0.2110

a. Write the multiple regression equation.

Price = 322.8 + 0.873(age) – 93.41(bidders) + 1.2979(age*bidders)

b. Test if the "Age" coefficient is significantly different from 0 and interpret the results.

   t = .873/2.020 = .432

   df = n – (k + 1) = 32 – (3 + 1) = 28

   p(.432, df = 28)

   p = .335(2) = .670. Therefore, we cannot reject the null hypothesis: age alone does not appear to have a significant effect on the selling price of a grandfather clock.

c. Construct a 92% confidence interval around the interaction coefficient and interpret.

   92% confidence ➔ t-value at .04 with 28 df = 1.817

   1.2979 ± 1.817(.2110) = 1.2979 ± .383 = [.9149, 1.681]. Because this interval does not include 0, we can say at a 92% confidence level that the interaction between age and bidders does have an effect on the price of a grandfather clock.

d. Draw the regression line that would result in a hypothetical situation with 0 bidders. Hypothetically, what does the y-intercept on this graph represent?

   With bidders = 0, the regression equation becomes y = 322.8 + .873(age)
   This would be a line with a y-intercept of 322.8 and a slight positive slope. Hypothetically, the y-intercept represents the price of a brand new (age = 0) grandfather clock with 0 bidders.

e. If the age of a clock is 100 years, what happens as you increase the number of bidders?

Price = 322.8 + 0.873(age) – 93.41(bidders) + 1.2979(age*bidders)

Price = 322.8 + 0.873(100) – 93.41(5, for example) + 1.2979(100*5) = $592

Price = 322.8 + 0.873(100) – 93.41(10, for example) + 1.2979(100*10) = $773.90

At age = 100 years, the price of a clock increases as the number of bidders increases.

f. If the age of a clock is 20 years, what happens as you increase the number of bidders? Compare this answer to part e—why does this happen? What does this mean about the regression equation?

Price = 322.8 + 0.873(age) – 93.41(bidders) + 1.2979(age*bidders)

Price = 322.8 + 0.873(20) – 93.41(5, for example) + 1.2979(20*5) = $3

Price = 322.8 + 0.873(20) – 93.41(10, for example) + 1.2979(20*10) = $-334.26

At age = 20, the price decreases with more bidders. This is because the coefficient for number of bidders is so large—if the age isn't high enough to cancel out the negative effect of the number of bidders in combination with the interaction term, the regression equation performs illogically. This regression equation is only meant to work on antique clocks (nobody really makes grandfather clocks anymore—so until the age of the clock is high enough, the regression equation does not work.

12. Which of the following are measures of spread?
    a. Median
    b. Standard deviation
    c. Range
    d. Root mean squared error

13. X is a random variable with mean μ and variance σ^2. The "standardized form" of X is Z = (X- μ)/ σ. What is the mean and variance, respectively, of Z?
    a. Mean = 0, variance = 1
    b. Mean = 1, variance = 0
    c. Mean = 0.05, variance = 0

14. A correlation coefficient of $r = 0.8$ is reported for a sample of pairs (x,y). Without any further information, this implies that…
    a. As the x values decrease, the y values increase.
    b. (x,y) are scattered about a straight line of unknown positive slope.
    c. 80% of the variation in y is due to regression on x.
    d. The points (x,y) are scattered about a straight line of slope 0.8.