Practice Final Exam Problems - STAT 100B

This practice exam includes all of your previous quiz questions as well as some additional practice problems. Questions from your homeworks are also good exam practice but, since you have open access to your textbook, those problems are not included here. The last four pages contain the formulas and tables that will be provided with your final exam.

Some study tips: I recommend mimicking exam conditions as much as possible as you study. Put away any distractions, find a quiet place to work, and do as much as possible using only a calculator and the formulas/tables provided at the end of this practice exam. The final exam will be approximately six questions (with multiple parts) and designed to take around 90 minutes, although you are more than welcome to take the full three hours. If you are worried about time, give yourself about 30 minutes per question. When you are done, return to your notes, textbook, or previous quizzes to check your work and figure out what your sticking points are. Use this information to guide your studying.

Previous Quiz Questions:

1. In a regression, what is the purpose of testing whether $\beta_1 = 0$? If we reject $H_0 : \beta_1 = 0$, does this imply a good fit? Explain.

2. Briefly explain the difference between confidence intervals and prediction intervals. Which interval would you expect to be wider?

3. Leonardo da Vinci (1452–1519) drew a sketch of a man, indicating that a person's arm span (measuring across the back with your arms outstretched to make a "T") is roughly equal to the person's height. To test this claim, we measured eight people with the following results:

Person	1	2	3	4	5	6	7	8
Arm span (in)	68	62.25	65	69.5	68	69	62	60.25
Height (in)	69	62	65	70	67	67	63	62

(a) Draw a scatterplot for arm span and height. Use the same scale on both the horizontal and vertical axes. Describe the relationship between the two variables.

(b) If da Vinci is correct, and a person's arm span is roughly the same as the person's height, what should the slope of the regression line be?

(c) The linear regression line for these data is `height` $= 12.22 + 0.82 \times$ `armspan`. Predict the height of a person who has a 66 inch armspan. Do you have any concerns about this prediction?

4. A shoe store has developed the following estimated regression equation relating sales to inventory investment and advertising expenditures:

$$\hat{y} = 25 + 10x_1 + 8x_2$$

where

$$x_1 = \text{inventory investment}$$
$$x_2 = \text{advertising expenditures}$$
$$y = \text{sales}$$

(a) Interpret b_1 and b_2 in this estimated regression equation.

(b) Estimate y when $x_1 = 180$ and $x_2 = 310$.

(c) For this model, $R^2 = 0.8472$ and $R^2_{adj} = 0.7923$. Comment on the goodness of fit.

5. Consider a regression study involving a dependent variable y, a quantitative predictor x_1, and a qualitative predictor variable with three levels (levels A, B, and C).

   (a) How many indicator (dummy) variables are required to represent the qualitative variable?

   (b) Write a multiple regression equation relating x_1 and the qualitative variable to y. Define (explain) any indicator variables. (Hint: use βs in your equation.)

   (c) Interpret the parameters in your regression equation.

Additional Practice Questions:

6. Seventy-six Starbucks food items were analyzed for the calorie and carbohydrate content. We used linear regression to explore the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain. The estimated regression equation with carbohydrates as the response variable and the calories as the explanatory variable is $\hat{y} = 8.94 + 0.11x$, and summary statistics of the two variables is provided below.

variable	min	Q1	median	Q3	max	mean	sd	n	missing
calories	80	300	350	420	500	338.8	105.4	77	0
carbohydrates	16	31	45	59	80	44.9	16.6	77	0

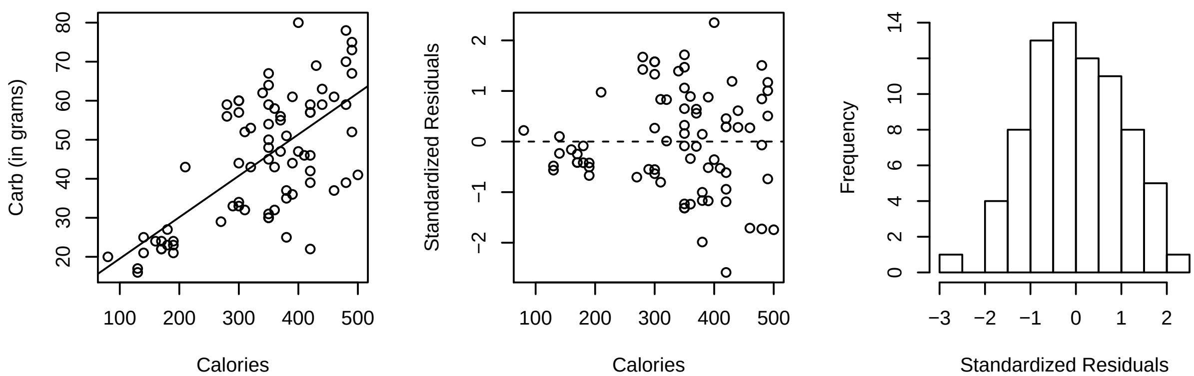
(a) Interpret the slope of the regression.

(b) Interpret the intercept of the regression.

(c) What is the predicted value for a food item that contains 300 calories and 50 grams of carbohydrates? Is this an over- or under-estimation? By how much?

(d) If the MSE is 10.1, calculate a 95% prediction interval for a food item that contains 300 calories and 50 grams of carbohydrates.

(e) The figures below show diagnostic plots for the regression. Are there any issues with the regression assumptions? If so, what assumptions are violated?



7. A multiple linear regression model is used to predict the mid-upper arm circumference (MUAC) of Ethiopian teenagers. The explanatory variables in the model are height (cm) and household income, measured in Ethiopian currency (the birr).

```
Residuals:
   Min     1Q Median     3Q    Max
-5.188 -1.851 -0.012  1.581  5.796

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.884958   5.713208   1.205  0.23420
height_cm    0.101084   0.036811   2.746  0.00852
hhincome    -0.000730   0.001388  -0.526  0.60142
---
Residual standard error: 2.459 on 47 degrees of freedom
Multiple R-squared:  0.1397,Adjusted R-squared:  0.1031
F-statistic: 3.817 on 2 and 47 DF,  p-value: 0.0291
```

(a) What is the estimated linear regression equation?

(b) Interpret the coefficient for `height_cm`.

(c) What is the null hypothesis being tested on the line where the p-value is 0.00852?

(d) Calculate a 95% confidence interval for the intercept.

8. Suppose we were to run an experiment where 24 bean plants are randomized into one of four groups:

   - Each plant receives 1 teaspoon of water and 1 hour of sunlight each day.
   - Each plant receives 4 tablespoons of water and 1 hour of sunlight each day.
   - Each plant receives 1 teaspoon of water and 8 hours of sunlight each day.
   - Each plant receives 4 tablespoons of water and 8 hours of sunlight each day.

   (a) Which group do you think will have the least plant growth?

   (b) The most plant growth?

   (c) How confident are you in your answers?

   (d) Do you think the effects of the water and sunlight on plans are independent? If so, explain why. If not, explain how you might model this relationship.

9. A poll asked 1253 adults a series of questions about the state of the economy and their children's future. One question was, "Do you expect your children to have a better life than you have had, a worse life, or a life about as good as yours?" The response breakdown was 34% "better", 29% "worse", 33% "about the same", and 4% "unsure". Use a nonparametric test at the 0.05 level of significance to determine if more adults feel their children will have a better future than feel their children will have a worse future. What is your conclusion?

10. Shown below are the number of baggage-related complaints per 1000 passengers for 10 airlines during the months of December 1988 and January 1989. Use a nonparametric test at the 0.05 level of significance to determine if the data indicate the number of baggage-related complaints for the airline industry has changed over the two months studied. What is your conclusion?

Airline	**December Complaints**	**January Complaints**
American	8.9	8.0
Delta	8.2	7.9
Continental	7.9	8.2
Eastern	7.5	7.8
Northwest	9.6	6.5
Pan American	5.0	5.1
Piedmont	12.3	11.0
TWA	11.2	10.9
United	7.7	7.4
USAir	8.6	7.9

Correlation

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\times\frac{y_i-\bar{y}}{s_y}\right)$$

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

$$R^2_{adj} = 1 - \frac{SS_{residuals}/(n-k-1)}{SS_{total}/(n-1)}$$

Least Squares Regression

$$b_1 = \frac{s_y}{s_x}R$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Logit Transformation

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Confidence/Prediction Intervals

$$\textit{point estimate} \pm (\textit{critical value}) \times (\textit{standard error})$$

Regression Confidence Intervals

$$SE(\hat{y}) = \sqrt{MSE\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_x}\right)}$$

under a $t(n-k-1)$ distribution.

Regression Prediction Intervals

$$SE(\hat{y}) = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_x}\right)}$$

under a $t(n-k-1)$ distribution.

Hypothesis Tests

$$\textit{test statistic} = \frac{\textit{point estimate} - \textit{null value}}{\textit{standard error}}$$

Normal Approximation to the Binomial Distribution

$$N\left(\mu = np, \sigma = \sqrt{np(1-p)}\right)$$

Case	**Test Statistic**	**Standard Error**
Paired Sample Means	$\frac{\bar{x}_{\text{diff}}-\mu_{\text{diff}}}{SE}$	$\frac{s}{\sqrt{n_{\text{pairs}}}}$
Independent Sample Means	$\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{SE}$	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
One Sample Median (large sample)	$\frac{(\#\text{ of + signs})-\mu}{SE}$	σ
Paired Samples (nonparametric)	$\frac{(\text{sum of signed ranks})-\mu}{SE}$	$\sqrt{\frac{n(n+1)(2n+1)}{6}}$

Critical Values for z

$(1-\alpha)100\%$	90%	95%	98%	99%
$z_{\alpha/2}$	1.645	1.96	2.33	2.575

Critical Values for t: $t_{\alpha/2,(n-1)}$

	$(1-\alpha)100\%$			
$(n-1)$	90%	95%	98%	99%
1	6.3137	12.706	31.821	63.657
2	2.9200	4.3026	6.9646	9.9248
3	2.3534	3.1824	4.5407	5.8409
4	2.1319	2.7765	3.7470	4.6041
5	2.0151	2.5706	3.3649	4.0321
6	1.9432	2.4469	3.1427	3.7074
7	1.8946	2.3646	2.9979	3.4995
8	1.8596	2.3060	2.8965	3.3554
9	1.8331	2.2622	2.8214	3.2498
10	1.8125	2.2281	2.7638	3.1693
11	1.7959	2.2010	2.7181	3.1058
12	1.7823	2.1788	2.6810	3.0545
13	1.7709	2.1604	2.6503	3.0123
14	1.7613	2.1448	2.6245	2.9768
15	1.7530	2.1315	2.6025	2.9467
16	1.7459	2.1199	2.5835	2.9208
17	1.7396	2.1098	2.5669	2.8982
18	1.7341	2.1009	2.5524	2.8784
19	1.7291	2.0930	2.5395	2.8609
20	1.7247	2.0860	2.5280	2.8453
21	1.7207	2.0796	2.5177	2.8314
22	1.7171	2.0739	2.5083	2.8188
23	1.7139	2.0687	2.4999	2.8073
24	1.7109	2.0639	2.4922	2.7969
25	1.7081	2.0595	2.4851	2.7874
26	1.7056	2.0555	2.4786	2.7787
27	1.7033	2.0518	2.4727	2.7707
28	1.7011	2.0484	2.4671	2.7633
29	1.6991	2.0452	2.4620	2.7564
30	1.6973	2.0423	2.4573	2.7500