Practice Exam #2 (Sta-209, F26)

Ryan Miller

Directions

- Answer each question using no more than the specified number of sentences and do not attempt to avoid these guidelines by using run-on sentences. Answers that are unnecessarily verbose may result in point loss.
- Do not include superfluous information in your answers, you may be penalized if you make an inaccurate statement even if you go on to provide a correct answer. Your answers should be clear, concise, and include only what is needed to answer the question that was asked.

Formulas and Definitions

Below are standard error formulas and large sample conditions used in hypothesis testing:

Population Parameter	Sample Statistic	Standard Error	Conditions
p	$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$np \geq 10$ and $n(1-p) \geq 10$
μ	$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	normal population or $n \geq 30$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i \in \{1,2\}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	normal populations or $n_1 \geq 30$ and $n_2 \geq 30$
β_j	$\hat{\beta}_j$	given in R output	independent, normally distributed residuals with constant variance

Below are a few definitions you don't need to memorize:

F-Test:

- SST = sum of squared residuals for null model
- SSE = sum of squared residuals for the alternative model
- d_0 = number of parameters in null model, d_1 = number of parameters in alternative model
- F-statistic = $\frac{(SST-SSE)/(d_1-d_0)}{SSE/(n-d_1)}$

Section 1 - True/False

Directions: Clearly indicate whether each of the following statements is true or false. You do not need to explain your reasoning and you will not receive a better score for doing so.

1. Testing $H_0 : \beta_1 = 0$ in a simple linear regression model is equivalent to testing the null hypothesis that Pearson's correlation equals zero.
2. If a simple linear regression model has a high R^2 value, such as $R^2 = 0.9$, Pearson's correlation will also be large in magnitude (close to either -1 or +1).
3. A benefit of simple linear regression is that it can be used to reliably make predictions about what is likely to occur outside of the range of the observed data, while Pearson's correlation only describes the relationship found within the range of the observed sample data.
4. In simple linear regression, an outlier must be extreme in both X and Y for it to exert a disproportionate impact on the slope coefficient of the fitted model.
5. Spearman's correlation is an effective way to measure the strength of association when a scatterplot shows a clear quadratic curve (U-shaped relationship).
6. In a regression model involving a single categorical predictor, the T-test produced by the `summary()` function for the *model's intercept* tests the null hypothesis that the mean of the reference category is equal to zero.
7. If side-by-side boxplots show a lot of within group variability and not much between group variability, one-way ANOVA is likely to produce a small p-value.
8. In a one-way ANOVA analysis, the alternative model uses group-specific means to calculate residuals, while the null model uses a single, overall mean to calculate residuals.
9. The slope coefficients in a regression model depend upon the units of the explanatory and response variables, so interpreting the expected effect of a 1-unit change in a predictor might not be appropriate in certain contexts.
10. The marginal effect of a predictor on an outcome can be estimated using a multivariable regression model whose predictors were chosen to provide the best fit to the sample data.
11. The fit of the models `Y ~ X1 + X2 + X3` and `Y ~ X1 + X4` can be compared using an F-test because the second model contains fewer predictors than the first.
12. In a regression analysis, it is possible that the marginal effect of a predictor is not statistically significant, but after adjusting for another variable the adjusted effect of that predictor is statistically significant.

Section 2 - Conceptual Questions

Directions: Answer each question using no more than 3-sentences. Do not include unnecessary details, as you will be penalized for any inaccurate statements, regardless of whether they are relevant or not. Aim to clearly answer the question that was posed, not to demonstrate your knowledge of related topics.

1. In your own words, briefly describe when Spearman's correlation should be used instead of Pearson's correlation in a correlation analysis. You should be specific in addressing what you'd want to look at to make the decision.

2. In Lab 12 you worked with the Iowa City Home Sales dataset, finding that the *marginal effect* of the variable "style" was such that 2-story homes were expected to sell for an average of $49,000 more than 1-story homes. However, after using regression to adjust for differences in "area.living", the *adjusted effect* was such that the coefficient of the variable "style" was now -41,000. In your own words, briefly explain why the marginal effect was such a large positive value, but the adjusted effect was much smaller.

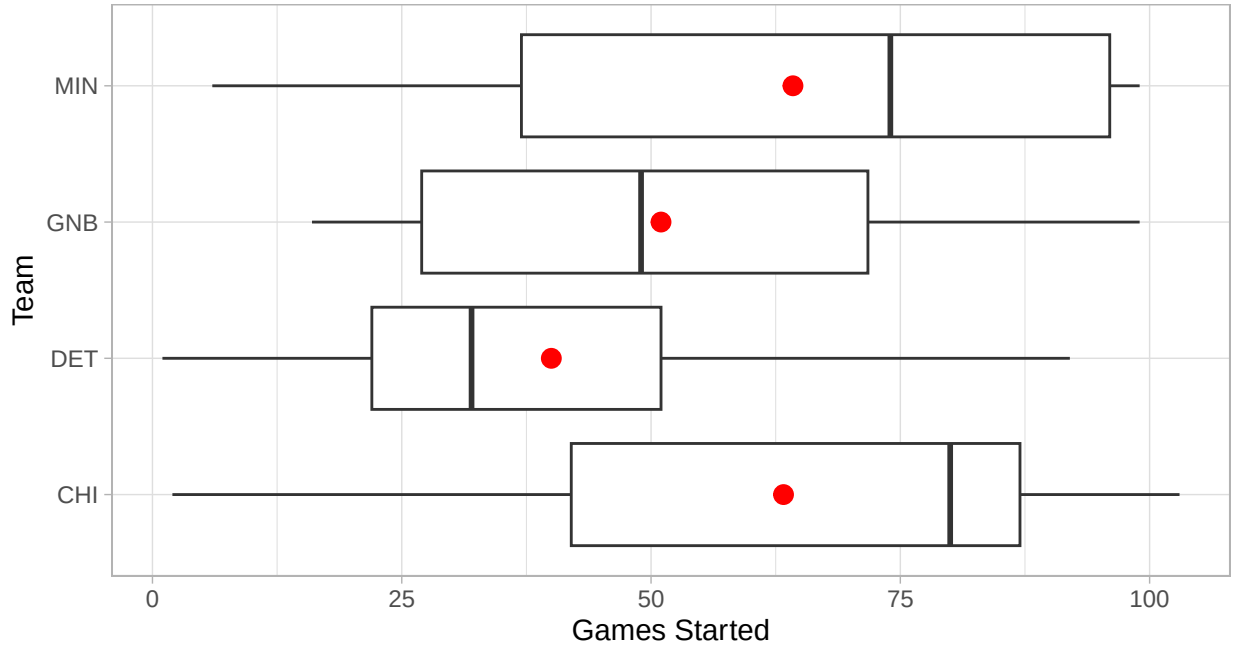
3. In your own words, explain how one-hot encoding will represent a categorical variable containing four categories: A, B, C, and D so that this variable can be used in a multivariable regression model containing it and a quantitative predictor.

4. In your own words, explain the purpose of *post hoc testing* (such as the Tukey HSD procedure or Dunn test) in one-way ANOVA. That is, when and why is post hoc testing used? What does it add to the ANOVA analysis?

Section 3 - Application #1

This application uses the NFL player value dataset first introduced in our correlation lecture. These data contain all players drafted in 2022. The data have been filtered to only include players from four NFL teams: Chicago, Detroit, Green Bay, and Minnesota.

Part A: Consider the boxplots given below showing the number of games started by players belonging to each of these four teams shown below. In this visualization, the mean of a team is shown by a dot. Based upon this visualization, do you believe it is reasonable to proceed in analyzing these data using a one-way ANOVA model?

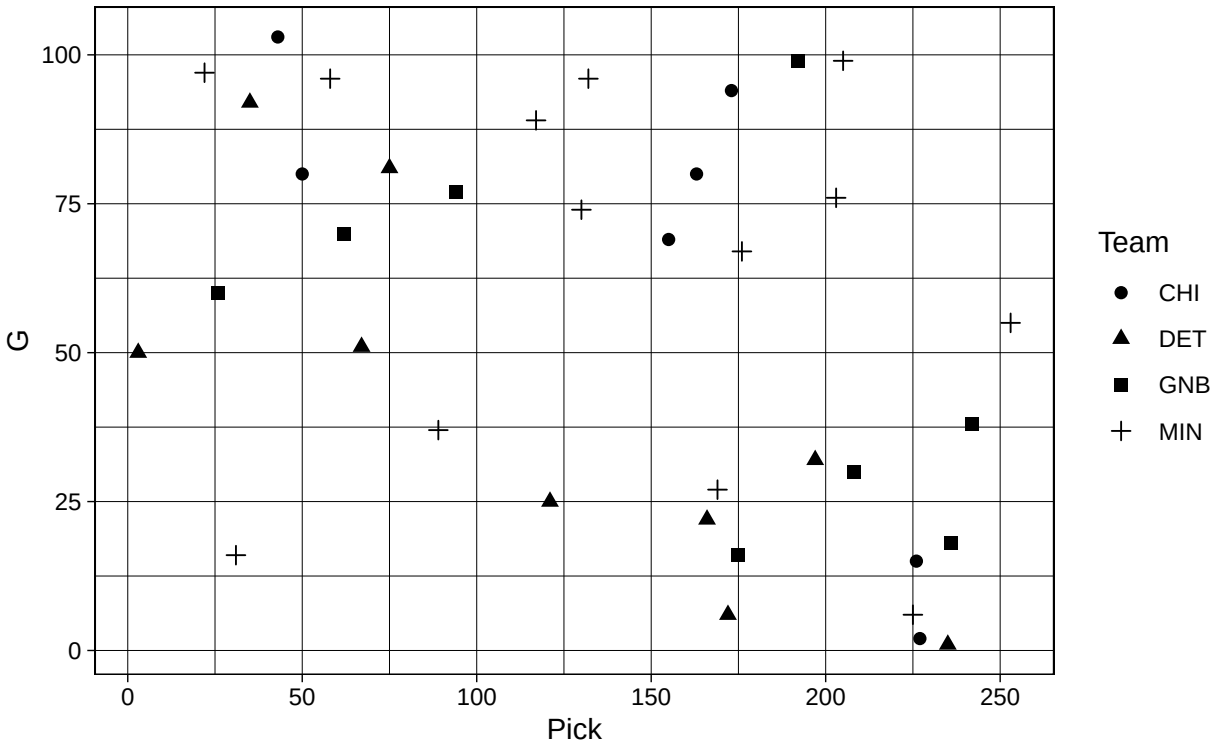


Part B: Shown below is the ANOVA table for the one-way ANOVA model using team to predict the number of games started. Provide a one-sentence interpretation of the results of this test.

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Tm           3   3738    1246   1.126  0.353
## Residuals   33  36528    1107
## 3 observations deleted due to missingness
```

Part C: Based upon the ANOVA table from Part B and the data visualization from Part A, does there appear to be more between group variation or within group variation in the number of games started for players in this dataset? Briefly explain your reasoning.

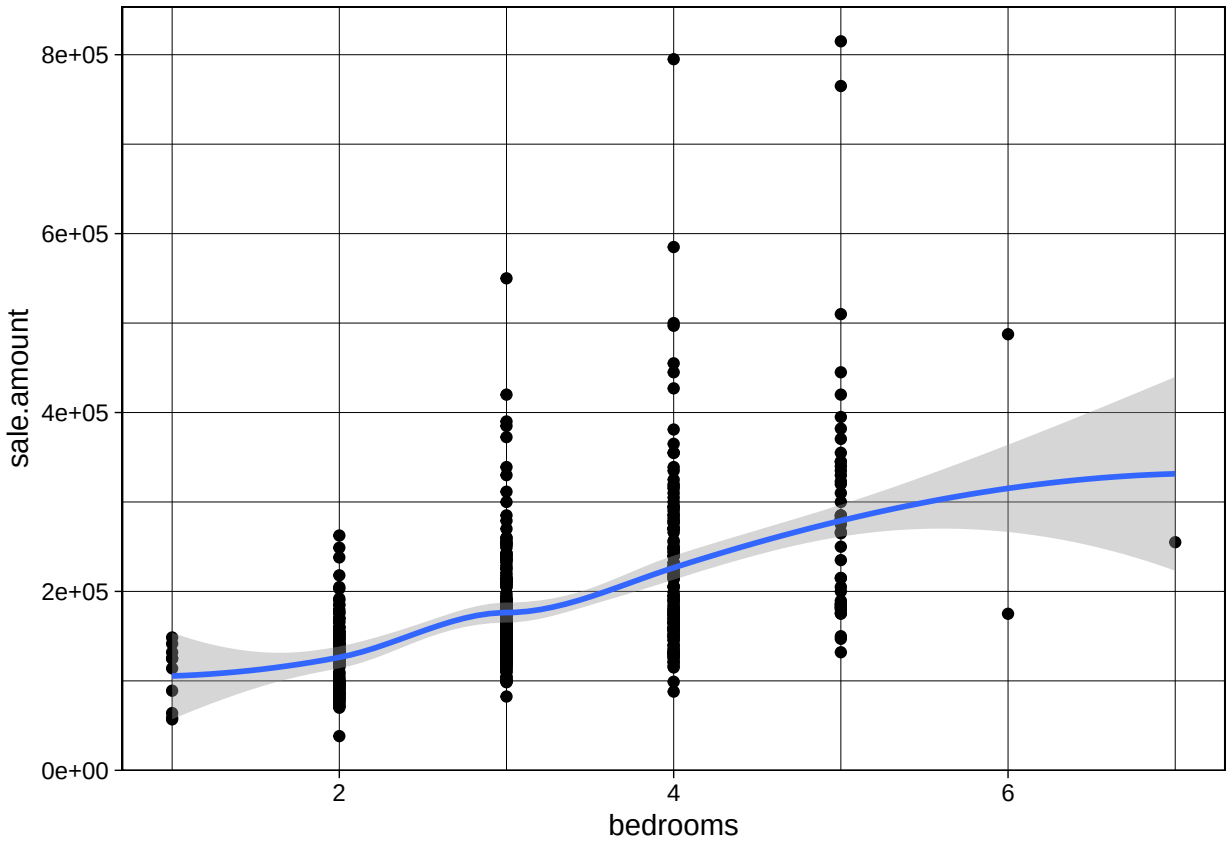
Part D: Suppose we expand the one-way ANOVA model by also using the quantitative predictor `Pick` which represents the pick in the NFL draft where the player was selected (note that earlier picks are made first, and thus tend to be better players). Sketch the *prediction line(s)* of this model, which uses the `R` formula `G ~ Team + Pick`, on the scatter plot below:


100
75
50
25
0
G
0
50
100
150
200
250
Pick
Team
CHI
DET
GNB
MIN

Section 3 - Application #2

In Lab 12 you worked with the Iowa City Home Sales dataset, which contains information on all single-family homes sold in Iowa City, IA in 2005-2008.

Part A: Shown below is a scatterplot showing the number of bedrooms in a home and its sale price with a loess smoother. Use this plot to qualitatively describe the relationship between bedrooms and sale price.



Part B: Consider a simple linear regression using the number of bedrooms a home has to predict its sale price. Based upon the `R` output below, provide a one-sentence conclusion for the T-test of $H_0 : \beta_1 = 0$ for this model. Be sure to include the strength of evidence, context, and the size of the observed effect (if one was found).

```
beds_model = lm(sale.amount ~ bedrooms, data = ic_homes)
summary(beds_model)
```

```
##
## Call:
## lm(formula = sale.amount ~ bedrooms, data = ic_homes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -156593  -44515   -8199   27832  565669
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    24805      10982   2.259   0.0243 *
```

```
## bedrooms       51131       3375  15.150   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77850 on 529 degrees of freedom
## Multiple R-squared:  0.3026, Adjusted R-squared:  0.3013
## F-statistic: 229.5 on 1 and 529 DF,  p-value: < 2.2e-16
```

Part C: Based upon the R^2 value of the model from Part A, does the number of bedrooms explain a lot of variation in sale price, a moderate amount of variation in sale price, or little variation in sale price?

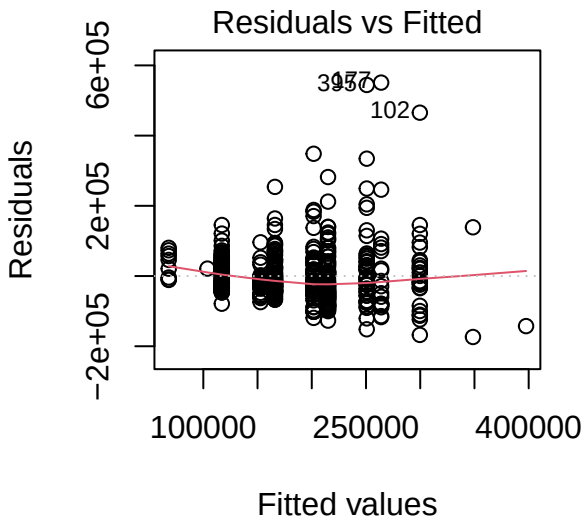
Part D: Now consider a second predictor, the style of the home (either 1-story or 2-story). On average, 2-story homes have more bedrooms than 1-story homes, meaning the likelihood of a home having the 2-story style increases as the number of bedrooms increases. Additionally, 2-story homes tend to be cheaper to build (on a per bedroom basis). Based upon this information, do you expect the effect of bedrooms on sale price to increase or decrease in a model that adjusts for style? Briefly explain your reasoning.

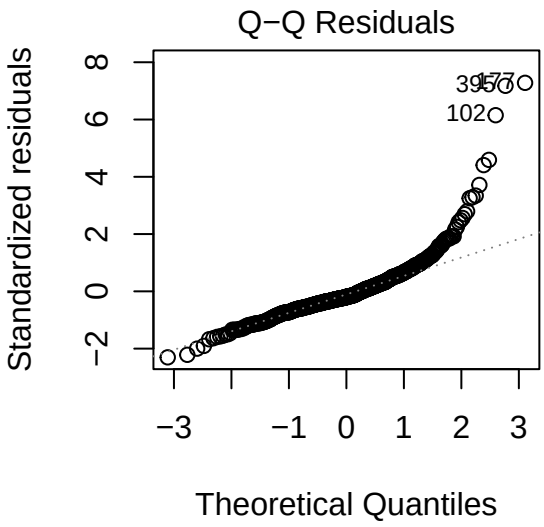
Part E: Suppose we use an F-test to compare the fit of the two-predictor model from Part D with the fit of the simple linear regression from Part B. Briefly interpret the results of this test (shown below, with the associated `R` code).

```
beds_model = lm(sale.amount ~ bedrooms, data = ic_homes)
beds_style_model = lm(sale.amount ~ bedrooms + style, data = ic_homes)
anova(beds_model, beds_style_model)
```

```
## Analysis of Variance Table
##
## Model 1: sale.amount ~ bedrooms
## Model 2: sale.amount ~ bedrooms + style
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    529 3.2064e+12
## 2    528 3.0557e+12  1 1.507e+11 26.039 4.675e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part F: Shown below are residual diagnostic plots for the two-predictor model from Parts D/E, `beds_style_model`. Based upon what you see in these plots, do you believe the assumptions for statistical inference using this model are satisfied? Briefly explain.


Residuals vs Fitted
Residuals
Fitted values
−2e+05
2e+05
6e+05
100000
250000
400000
102


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
−2
0
2
4
6
8
−3
−1
0
1
2
3
102