

Practice Exam #3 - Sketch Solution (Sta-209, S25)

Ryan Miller

The following information will appear verbatim on the first page of Exam 3.

Directions

- Answer each question using *no more than specified number of sentences* and 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.

Formula Sheet

Theoretical models:

- Linear regression: $y_i = \beta_0 + \beta_1 \cdot x_{i1} + \dots + \beta_p \cdot x_{ip} + \epsilon_i$
- Logistic regression $\log\left(\frac{Pr(y=1)}{1-Pr(y=1)}\right) = \beta_0 + \beta_1 \cdot x_{i1} + \dots + \beta_p \cdot x_{ip}$

Chi-squared test statistic:

$$X^2 = \sum_{j=1}^k \frac{(\text{observed}_j - \text{expected}_j)^2}{\text{expected}_j}$$

F-test statistic:

$$F = \frac{(SS_0 - SS_1)/(d_1 - d_0)}{SS_1/(n - d_1)}$$

Reminders:

- Odds - a ratio of how often an event occurs relative to how often it doesn't occur
- Linear regression assumptions - residuals are independent and Normally distributed with constant variance

Question #1 (conceptual questions)

Part A: Suppose we are interested in building a linear regression model that predicts daily ozone concentration based upon three quantitative explanatory variables: temperature, wind speed, and solar radiation. Identify which of the following statements must be **true** (there may be more than 1 true statement):

- A) The model: $\widehat{Ozone} = b_0 + b_1Temp + b_2Wind$ will have a smaller sum of squared residuals than the model $\widehat{Ozone} = b_0 + b_1Solar$
- *False* - these models aren't nested, and we don't know which explanatory variables are most predictive.
- B) The model: $\widehat{Ozone} = b_0 + b_1Temp + b_2Wind$ will have a smaller sum of squared residuals than the model $\widehat{Ozone} = b_0 + b_1Temp$
- *True* - these models are nested, so the larger model will improve the sum of squared residuals by some amount, though the improvement might be small if **Wind** is unrelated to **Ozone**
- C) The model: $\widehat{Ozone} = b_0 + b_1Temp + b_2Temp^2$ will have a smaller sum of squared residuals than the model $\widehat{Ozone} = b_0 + b_1Wind$
- *False* - these models aren't nested, and we don't know which explanatory variables are most predictive.

State which statements are true and briefly explain the reasoning or thought process you used to determine whether a statement was true or false.

Part B: For each of the following scenarios state the name of the appropriate statistical model/hypothesis test. You do not need to explain your answers.

- i: Using a sample Grinnell students from the science division to see if the racial/ethnic distribution of science students at Grinnell differs from the distribution of the entire student body that is published by the college.
 - Chi-squared goodness of fit test
- ii: Conducting a randomized experiment to determine whether fertilizer A, fertilizer B, or fertilizer C will have different crop yields.
 - One-way ANOVA
- iii: Understanding how highschool GPA, ACT score, and number of extracurricular activities relate to whether or not students are accepted into Ivy League schools.
 - Logistic regression
- iv: Evaluating whether homes are more likely to have solar panels in certain regions of the country than others.
 - Chi-squared test of independence.

Part C: Recall that one-way ANOVA can be described as a comparison between two models using the observed sample data. With this in mind, answer the following questions:

- i: Suppose we are interested in how each model involved in one-way ANOVA will *predict* the value of the outcome variable. Briefly describe what the prediction will be based upon for each model.
 - The null model will base its predictions on the *overall mean* of the entire sample, ignoring the grouping variable. The alternative model will base its predictions of *group-specific means*, where groups are defined by a categorical explanatory variable.
- ii: The models in one-way ANOVA involve the Normal distribution. Briefly describe the role of the Normal distribution in these models.
 - The models assume *Normally distributed errors*, which means that any deviations from the predicted values will follow a Normal distribution (ie: Normally distributed residuals)
- iii: Suppose we perform one-way ANOVA and reject the null hypothesis. We check the model's assumptions and they are verified as reasonable. Is this the end of our analysis or is there more that we

should do? If this is end, briefly describe what we'd conclude from the test (in generic terms). If more should be done, briefly describe what you'd do next.

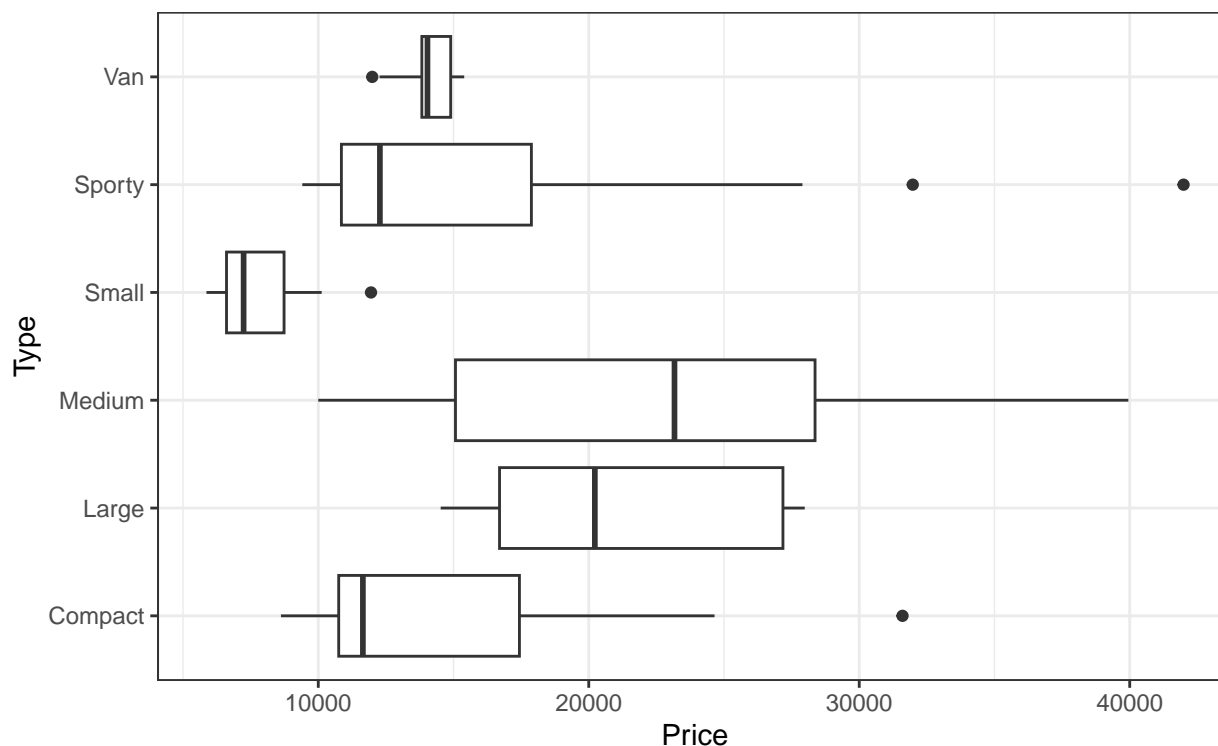
- We should conclude that there is an association between the explanatory and response variables, which implies that at least one group has a different mean. We should follow this up by performing post-hoc testing to determine which pairings of groups have different means.

Question #2

This question will analyze data on 111 different types of cars published in *Consumer Reports*. The overall goal of the analysis is to identify factors associated with price. A few key variables include:

- **Price** - List price (US dollars) with standard equipment
- **Country** - Where the car was manufactured
- **HP** - Net horsepower
- **Type** - A categorical variable describing the general type of vehicle (small, medium, large, compact, sporty, van)
- **Length** - Length of the vehicle (inches)

Part A: The plot below shows the relationship between **Type** and **Price**. Based upon the plot, do you believe these variables are associated? What is the name of the statistical test you'd use to determine whether or not there is evidence of an association?



- Yes, they are associated due to the large differences in median price. One-way ANOVA is the appropriate statistical test.

Part B: The table below summarizes price by vehicle type. Is any information presented in this table problematic for the validity of the statistical test you identified in Part A? If so, briefly explain what aspect(s) of these data are problematic.

Type	N	Mean	Median	StdDev
Compact	19	14395.37	11650.0	5938.76
Large	7	21499.71	20225.0	5825.88
Medium	26	22750.15	23170.0	8416.81
Small	22	7736.59	7239.5	1627.93
Sporty	21	15889.81	12279.0	8539.24
Van	10	14014.30	14037.5	1126.10

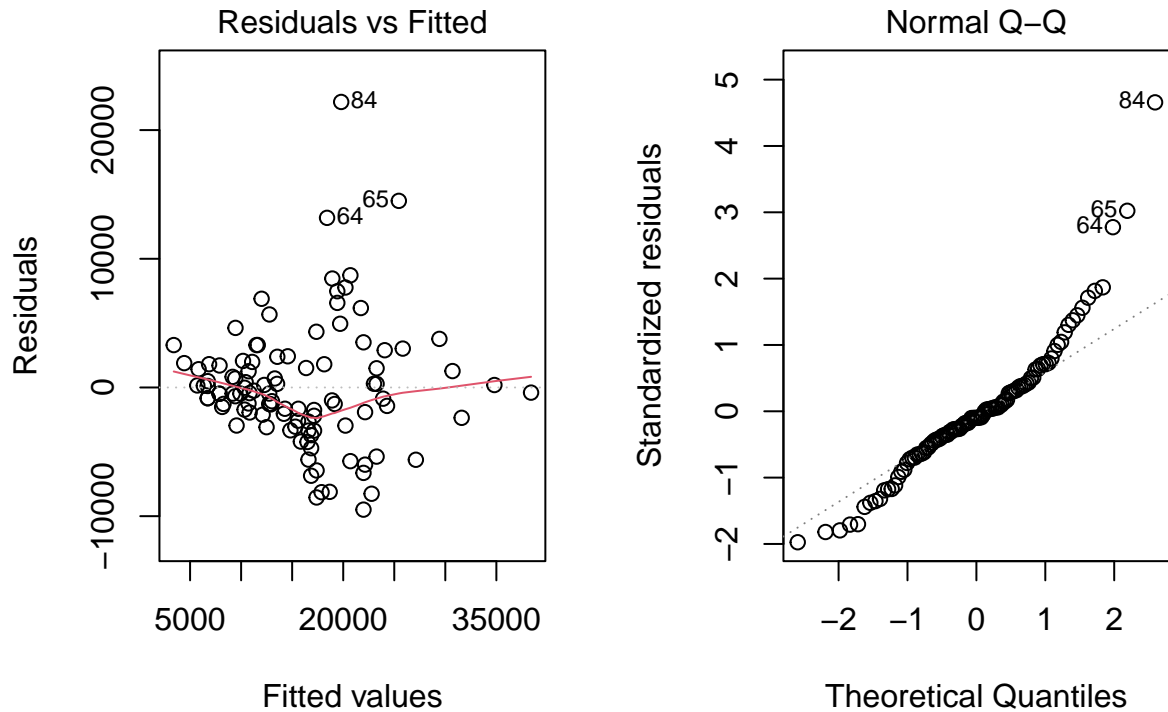
- The equal variances condition of ANOVA is not reasonable as some groups have standard deviations that are 7-8 times larger than other groups. Note that we cannot determine the Normality of residuals from this table, so we cannot make claims about the other main assumption of ANOVA.

Part C: The table below displays the *coefficient estimates* of a linear regression model that uses *both* Type and HP to predict a vehicle's Price. Use the information in this table to answer the following questions (I - III)

	Coefficient	Std. Error	t statistic	p-value
(Intercept)	-1842.01	2197.32	-0.84	0.40
HP	128.23	14.91	8.60	0.00
TypeLarge	2825.67	2224.20	1.27	0.21
TypeMedium	4593.94	1543.07	2.98	0.00
TypeSmall	-1810.14	1635.77	-1.11	0.27
TypeSporty	488.56	1556.85	0.31	0.75
TypeVan	-248.79	1915.62	-0.13	0.90

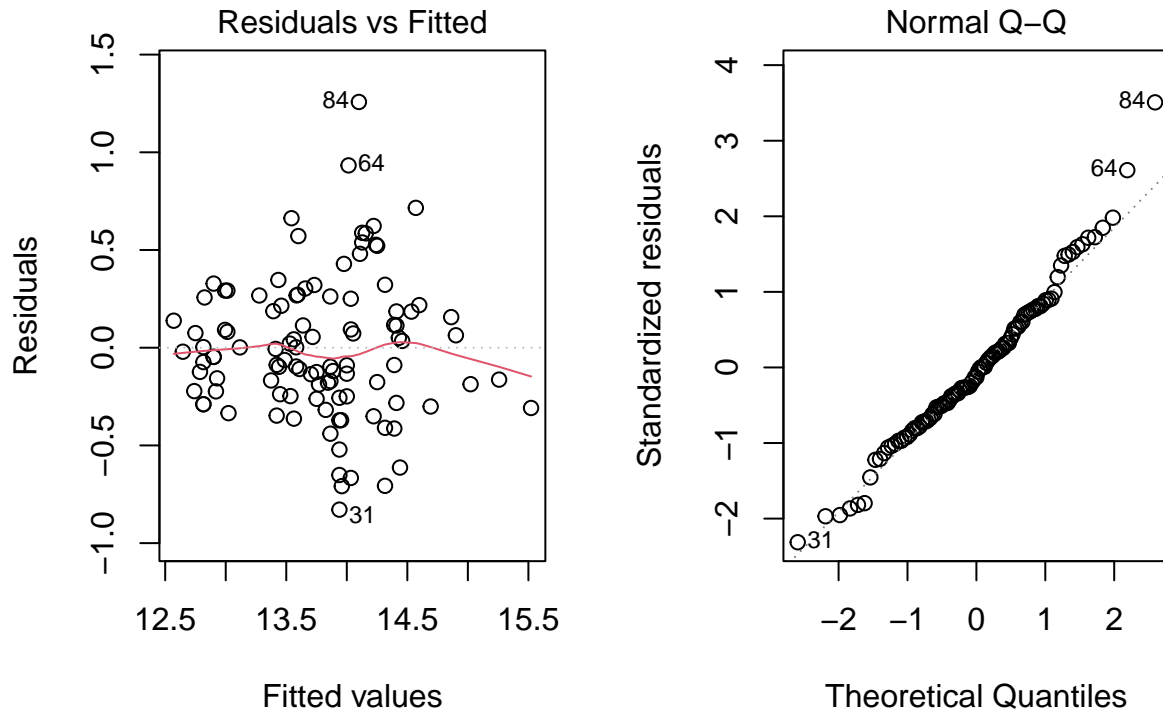
- I) The intercept of this model is -1842.01, what does this value mean? Should we care that this value isn't statistically significant?
 - For a compact car with zero horsepower the expected sale price is -1842.01 dollars. We shouldn't care about this value or its statistical significance since no cars have zero horsepower.
- II) Provide a one sentence interpretation of the coefficient for "TypeMedium", be specific.
 - A medium car is expected to sell for 4593.94 dollars more than a compact car after adjusting for differences in horsepower (ie: assuming both have the same horsepower)
- III) True or False, in this model the effect of HP on price differs depending on the type of vehicle. You do not need to explain your answer.
 - *False*, in a linear regression model each explanatory variable makes a separate additive contribution towards the outcome.

Part D: Below are two R plots related to the model described in Part D, $\text{Price} \sim \text{HP} + \text{Type}$. Based upon what you see in these plots, do you believe *p*-values calculated for these data will be valid/reliable? Briefly explain.



- No, there are several assumptions for inference that are violated. The first plot suggests the residuals do not have equal variance across all fitted values, and the second plot suggests the residuals are right-skewed and do not follow a Normal distribution.

Part E: The plots show results after transforming the response variable **Price** using a log-transformation, making the model: $\log_2(\text{Price}) \sim \text{HP} + \text{Type}$. When compared with the model from Parts D-E, are you *more comfortable* trusting the p -values produced by statistical tests that involve this model? Briefly explain why or why not.



- Yes, I am more comfortable now. While the constant variance assumption might still be questionable, it looks more reasonable than it did prior to the transformation. And the Normality assumption is now met.

Part F: Below are statistical results found using R. Based upon what is given, state the null hypothesis of the test that was performed in words and provide a one-sentence conclusion describing the results of the test in regard to the null hypothesis.

```
mod0 <- lm(log2(Price) ~ HP, data = car90)
mod1 <- lm(log2(Price) ~ HP + Type, data = car90)
anova(mod0, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: log2(Price) ~ HP
## Model 2: log2(Price) ~ HP + Type
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1     103 19.333
## 2      98 13.347   5    5.9861 8.7906 6.427e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The null hypothesis was that both models, mod0 and mod1, fit the data equally well. You could also state that the null hypothesis is that Type is independent of Price after adjusting for differences in HP. The hypothesis test provides strong evidence against this null hypothesis, suggesting that using Type and HP to predict price is better than just using HP.

Part G: Below are the estimated coefficients in `mod1` described in Part F. Interpret the coefficient of the variable `HP` in this model. Be careful to recognize that the outcome in the model has been log-transformed.

```
## (Intercept)          HP      TypeLarge  TypeMedium  TypeSmall  TypeSporty
## 12.525579939  0.009424088  0.311799577  0.376645754 -0.473314559  0.019619029
##      TypeVan
## 0.060987375
```

- We must undo the transformation on Price. Noting that $2^{0.00942} = 1.0065$, we conclude that for two vehicles of the same type (ie: controlling for Type), a 1 horsepower increase is expected to increase price by a multiplicative factor of 1.0065, or 0.65%

Question #3

The Donner party was a famous expedition of pioneers traveling to California through the Sierra Nevada mountains. The group became stranded in the mountains and spent the winter of 1846-1847 snowbound. Nearly one-half of the party starved to death before they were able to successfully escape the mountains.

Part A: Consider a Chi-squared test of independence involving the variables `sex` and `survival` using the table provided below. How many females would be expected to have died if `sex` and `survival` were independent?

```
##
##      died survived
## Female    10     25
## Male     32     24
```

- Overall 46.15% of the Donner party died, so if sex and survival were independent we'd expect 46.15% of the 35 females in the party to have died, or 16.15 females to have died.

Part B: Below are the results of the Chi-squared test mentioned in Part A. Briefly interpret these results in the context of the study and the test's hypotheses.

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(Donner$sex, Donner$y)
## X-squared = 5.9719, df = 1, p-value = 0.01454
```

- There is strong evidence of an association between sex and survival status for members of the Donner party.

Part C: Below is a logistic regression model using *both* `sex` and `age` to predict survival. Interpret the estimated intercept of this model.

```
##
## Call:  glm(formula = y_binary ~ age + sex, family = "binomial", data = Donner)
##
## Coefficients:
## (Intercept)          age      sexMale
##    1.62180    -0.03561   -1.06798
##
## Degrees of Freedom: 87 Total (i.e. Null);  85 Residual
```

```
## (3 observations deleted due to missingness)
## Null Deviance:      120.9
## Residual Deviance: 108.9    AIC: 114.9
```

- $\exp(1.62180) = 5.062$, this is the estimated odds of survival for a zero year-old female. So, we'd expect 5 newborn female babies to survive for every 1 that dies.

Part D: Interpret the estimated coefficient of **age** in the model from Part C.

- $\exp(-0.03561) = 0.965$, this is the multiplicative change in the odds of survival for each additional year in age, after adjusting for sex. So, for two individuals of the same sex the odds of survival decrease by 3.5% for each additional year in age.

Part E: Interpret the estimated coefficient of **sexMale** in the model from Part C.

- $\exp(-1.06798) = 0.343$, this is the odds ratio comparing the odds of survival for males with the odds of survival for females, after adjusting for differences in age. We can conclude that males have 65.7% lower odds of survival than females of the same age.

Part F: Below is a summary table for the model described in Part C. Based upon these results, do you believe that **sex** and **survival** are independent *after adjusting for differences in age*? Briefly explain.

```
##           Estimate Std. Error   z value   Pr(>|z|)
## (Intercept)  1.62180162 0.50279192  3.225592 0.001257124
## age         -0.03560529 0.01524512 -2.335521 0.019516239
## sexMale     -1.06797669 0.48228705 -2.214401 0.026801239
```

- No, there is strong evidence that the coefficient estimate for **sexMale** is not zero, suggesting that males are significantly less likely to survive even after adjusting for differences in age.