# Swarthmore Honors Exam 2015: Statistics

John W. Emerson, Yale University

**NAME:** _____

## Instructions:

This is a closed-book three-hour exam having 7 questions. You may not refer to notes or textbooks. You may use a calculator that does not do algebra or calculus. Normal, $t$, and $\chi^2$ tables should be supplied with this exam. Please note:

- The last two or three questions have a considerable amount of description and background which is intended to help you. Read this material carefully and completely before working on the problem.
- Some questions have multiple parts. Number the questions and parts clearly in your work, and start each of the questions on a new page.
- Questions that explicitly ask for (or imply the need for) discussion are a chance to demonstrate your understanding of the material. As a guideline, a short paragraph is likely appropriate and preferable to either a single sentence or a longer essay.
- The chi-squared density with $k$ degrees of freedom is $\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}$ for $x \geq 0$; I assume that you are familiar with the normal, exponential, and uniform density functions.
- The Poisson distribution has probability mass function $f_\lambda(k) = \lambda^k e^{-\lambda}/k!$ for $k = 0, 1, 2, \ldots$; I assume you are well-familiar with the Binomial distribution.

**1.** Random samples of 36 Williams College students and 49 Swarthmore College students are selected. Researchers discover that 18 of the Williams College students took AP Statistics in high school, while 14 or the Swarthmore College students had taken AP Statistics in high school. On the surface, the results (50% versus 28.6%, respectively) may seem surprisingly different, but the sample sizes are small. What do you think? Support your work with formal calculations and a concise presentation and defense of your choice of methodology.

**2.** Suppose $X_i \sim N(0,1)$ for $i = 1, 2$ independently. For each of the following problems, show your work. Pictures are encouraged. If a problem can't reasonably be calculated, explain why and provide the best partial solution that you can.

a. What is $P(X_1 > 1)$?

b. What is $P(|X_1| > 2)$?

c. What is $P(X^2 > 3)$?

d. What is $P(X_2 > 1 | X_1 > 1)$?

e. Let $Y_1 = |X_1| + |X_2|$. What is $P(Y_1 > 1)$?

f. Let $Y_2 = X_1^2 + X_2^2$. What is $P(Y_2 > 4)$?

g. Again, $Y_2 = X_1^2 + X_2^2$. What is $E(Y_2 | Y_2 > 4)$?

**3.** Suppose $X \sim Poisson(\lambda)$. Unfortunately, the events contributing to the count $X$ are only observed with some probability $p$ (independently), leading instead to a count of events $Y$. You could think about $X$ as the number of traffic accidents at a certain intersection in some interval of time, for example. Suppose that with probability $p$ a given accident results in the fire department being called to the scene, independently across accidents. Then $Y$ in this story would denote the number of accidents at the intersection resulting in the fire department being called to the scene. Find the distribution of $Y$; show your work.

**4.** Suppose $X_1$ and $X_2$ are independent $Uniform(\mu - 1/\mu, \mu + 1/\mu)$ for some $\mu > 0$. Derive the maximum likelihood and method of moments estimators.

**5.** You are asked to study professional basketball (the NBA) teams in a certain season, right before the playoffs. The model you will construct and could estimate (if you were actually doing this with real data and a computer) could be used to predict the outcome of games in the playoffs. You have the entire season of results right before the playoffs and are given a data set that looks like the following (but of course with many many more games, each in a single row of the data set):

```
> head(x, 3)

    team1   team2 scorediff location
1 Celtics    Nets         8        H
2   Bulls Celtics        -2        V
3 Pistons    Heat        13        V
```

In these examples, the Boston Celtics beat the Nets by 8 points playing at home in Boston; the Bulls went on the road to Boston and lost to the Celtics by 2 points, and the Pistons went on the road to Miami and beat the Heat by 13 points. A simple model is proposed that would have a coefficient representing the "strength" of each team. The idea is that the expected score differential would be the difference between two teams' strength coefficients. Ignoring the home court advantage, then, the simple model for a game between the Celtics and the Nets (with the Celtics as "team1") would be:

$$\text{scorediff}_{Celtics\ vs\ Nets} = \beta_{Celtics} - \beta_{Nets} + \varepsilon$$

where $\varepsilon$ is assumed to be approximately $N(0, \sigma^2)$. We would like positive strength coefficients to represent "above average strength" and negative coefficients to represent "below average strength", loosely speaking.

Show the construction of the model matrix corresponding to this linear model. Pay particular attention to the use (or omission) of an intercept and also provide for estimation of an overall "home team advantage" coefficient, $\alpha$. The interpretation of this coefficient would be that, on average, the home team enjoys an advantage of about $\alpha$ points above whatever is predicted by the difference in the strength coefficients. Unlike college basketball, there are no neutral-site games.

**Special Note:** Here you may assume that I've been careful with the data (unlike a similar Swarthmore problem were I intentionally didn't clean up the data). Specifically, each game appears once and only once in the data set. The order of listing of "team1" and "team2" is arbitrary on any given line and not important for this problem.

**6. 1.5-Carat diamonds and cut.** You should read the next few pages and look at the plots before answering the questions posed on page 8.

A diamond merchant is studying a great collection of high-quality 1.5-carat diamonds; these diamonds are better than "Fair" cut, so you won't find any "Fair" cut diamonds in this data set (even though "Fair" is a well-known cut in the diamond world). She also restricts her diamonds to having good (clear) coloring (essentially a lack of color). But unlike her competitors, she uses a new measurement technology to grade the color (`colorqual`) on a continuous scale (with smaller values being the poorer coloring and higher values being better – though none of these diamonds have very undesirable coloring values, which would be negative). But you won't consider clarity until Question 7.

The diamond merchant took *Introductory Statistics* in college and has even used **R** a little bit. She wants to explore the importance of the diamond cut on the price and does the following:

```
> x <- read.csv("diamonds_large.csv", as.is=TRUE)
> dim(x)

[1] 148   4

> head(x)

  price       cut colorqual clarity
1 12291 Very Good 0.7788804      IF
2  9424   Premium 3.5060028      IF
3 13075   Premium 5.5089217      IF
4 13107   Premium 4.5425640      IF
5 10332   Premium 5.3441561      IF
6 12437     Ideal 4.7291995      IF

> tail(x)

    price       cut colorqual clarity
143  8258   Premium 0.6433514     VS1
144  7847     Ideal 3.3201246     VS2
145 11748 Very Good 3.1471204    VVS2
146  7370      Good 1.5334632     VS2
147  8952 Very Good 6.2293858     VS2
148 10023 Very Good 6.0968066     VS2

> table(x$cut)

     Good     Ideal   Premium Very Good
       15        37        54        42
```

```
> x$cut.f <- factor(x$cut)
> lm.cut <- lm(price ~ cut.f, data=x)
> summary(lm.cut)

Call:
lm(formula = price ~ cut.f, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-3709.9 -1215.2   -10.3  1211.1  4252.2

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      8995.8      420.7  21.382   <2e-16 ***
cut.fIdeal       1132.9      498.8   2.271   0.0246 *
cut.fPremium     1056.1      475.6   2.221   0.0279 *
cut.fVery Good    721.6      490.1   1.472   0.1432
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 1629 on 144 degrees of freedom
Multiple R-squared:  0.04186,        Adjusted R-squared:  0.0219
F-statistic: 2.097 on 3 and 144 DF,  p-value: 0.1032

> anova(lm.cut)

Analysis of Variance Table

Response: price
           Df    Sum Sq Mean Sq F value Pr(>F)
cut.f       3  16705198 5568399  2.0973 0.1032
Residuals 144 382330715 2655074
```

She explores this basic association with side-by-side boxplots (next page, Figure 1), and then hides the real residual normal quantile plot on the subsequent page along with 8 simuated normal quantile plots (Figure 2, a neat trick to test your ability to interpret such a plot). She doesn't tell you what `pos` equals (you can probably deduce what is happening with the code)!

```
> boxplot(price ~ cut.f, data=x,
+         xlab="Cut of Diamond", ylab="Price of Diamond ($)")
```

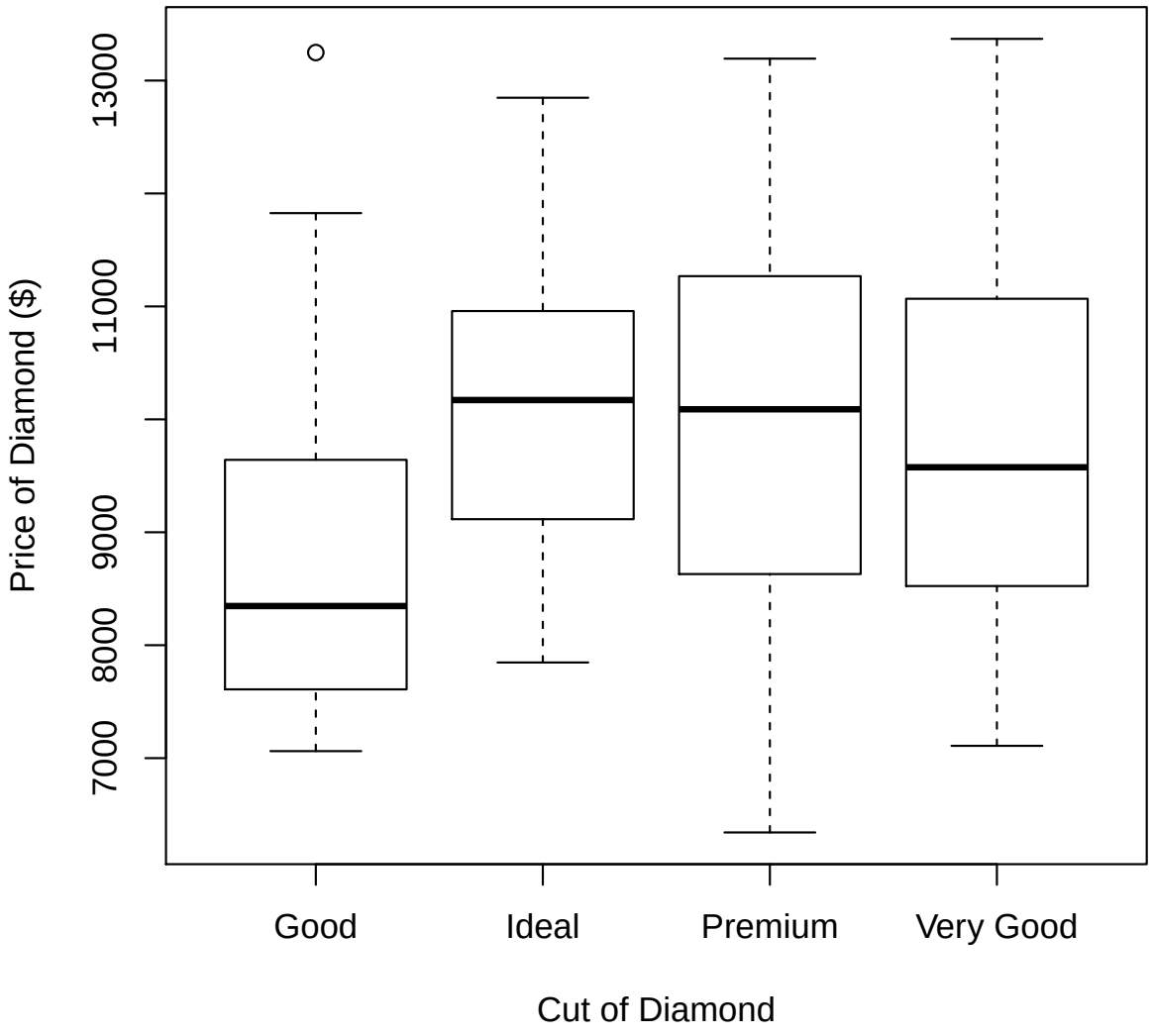


Figure 1: Boxplot of `price` by `cut.f`.

```
> par(mfrow=c(3,3), yaxt="n", xaxt="n")
> for (i in 1:9) {
+   if (i==pos) qqnorm(lm.cut$resid) else qqnorm(rnorm(nrow(x)))
+ }
```

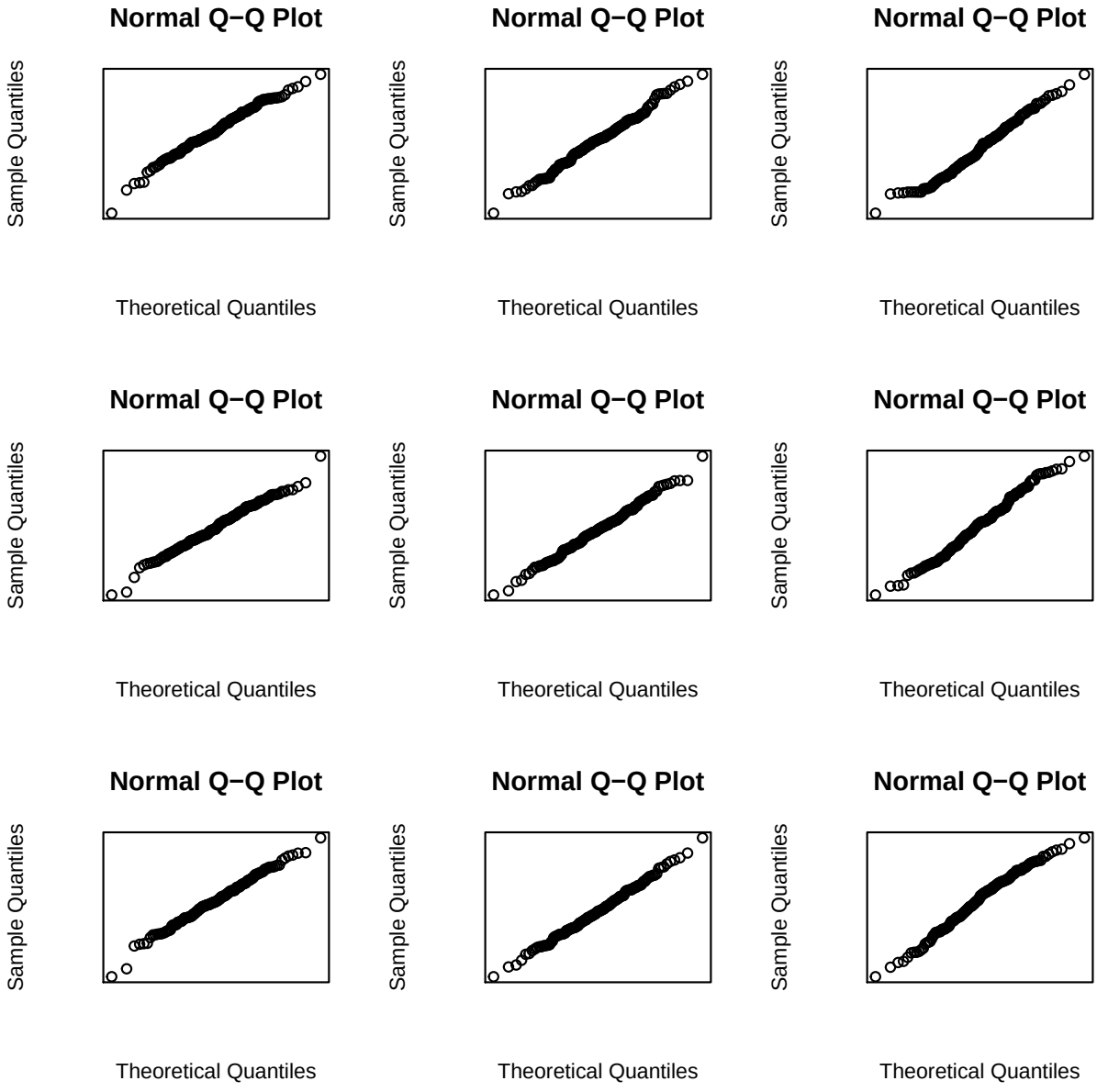


Figure 2: Some normal quantile plots.

## QUESTIONS ON THE MATERIAL APPEARING ABOVE

**Question 6A:** Based on this simple model, what is the predicted price of an "Ideal" cut diamond? Show the algebra – don't just give the answer.

**Question 6B:** A new 1.5-carat diamond is added to the collection, of "Ideal" cut and with a price tag of $7,000. Calculate its approximate residual (obviously you can't refit the model, but it wouldn't change much if you did). Again, show the algebra. Are you surprised by the price? Explain why or why not.

**Question 6C:** Based on the analysis presented on the previous pages, does the cut of the diamond appear to be a statistically significant predictor of price? Please reference any statistics you use to answer the question and explain how you determined "statistical significance" or lack thereof.

**Question 6D:** Are you concerned that something apparent in residual plots or the regression summary should be a serious cause for concern? If so, what, specifically, concerns you? If nothing concerns you, briefly indicate one characteristic of the residual plots or summaries that shows that you shouldn't be overly concerned.

**7. 1.5-Carat diamonds, multivariate.** Continue reading. The next questions appear on page 14.

The merchant then tries fitting models with all available predicting variables, but she is also concerned that perhaps the cut of the diamond may not be a statistically significant predicting variable. She does the following work.

```
> table(x$clarity)

  IF  VS1  VS2 VVS1 VVS2
  18   41   63   10   16

> x$clarity.f <- factor(x$clarity)
> lm.full <- lm(price ~ clarity.f + colorqual + cut.f, data=x)
> summary(lm.full)

Call:
lm(formula = price ~ clarity.f + colorqual + cut.f, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-2935.06  -901.76   -38.06   819.57  2365.44

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      9056.7      509.2  17.786  < 2e-16 ***
clarity.fVS1    -1794.8      350.1  -5.126 9.71e-07 ***
clarity.fVS2    -2483.2      327.1  -7.593 4.14e-12 ***
clarity.fVVS1   -1324.8      483.4  -2.741  0.00694 **
clarity.fVVS2   -1713.4      422.8  -4.052 8.40e-05 ***
colorqual         514.1       72.5   7.090 6.20e-11 ***
cut.fIdeal       1231.0      374.8   3.285  0.00129 **
cut.fPremium      856.1      358.7   2.387  0.01834 *
cut.fVery Good    777.2      371.8   2.090  0.03841 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 1219 on 139 degrees of freedom
Multiple R-squared:  0.4824,        Adjusted R-squared:  0.4526
F-statistic: 16.19 on 8 and 139 DF,  p-value: < 2.2e-16

> anova(lm.full)

Analysis of Variance Table

Response: price
```

```
            Df    Sum Sq  Mean Sq F value    Pr(>F)
clarity.f    4 105528592 26382148 17.7542 8.552e-12 ***
colorqual    1  70724108 70724108 47.5948 1.697e-10 ***
cut.f        3  16234268  5411423  3.6417    0.0144 *
Residuals  139 206548944  1485964
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

```
> par(mfrow=c(1,1), xaxt="s", yaxt="s")
> gpairs(x[,c("price", "cut.f", "clarity.f", "colorqual")])
```

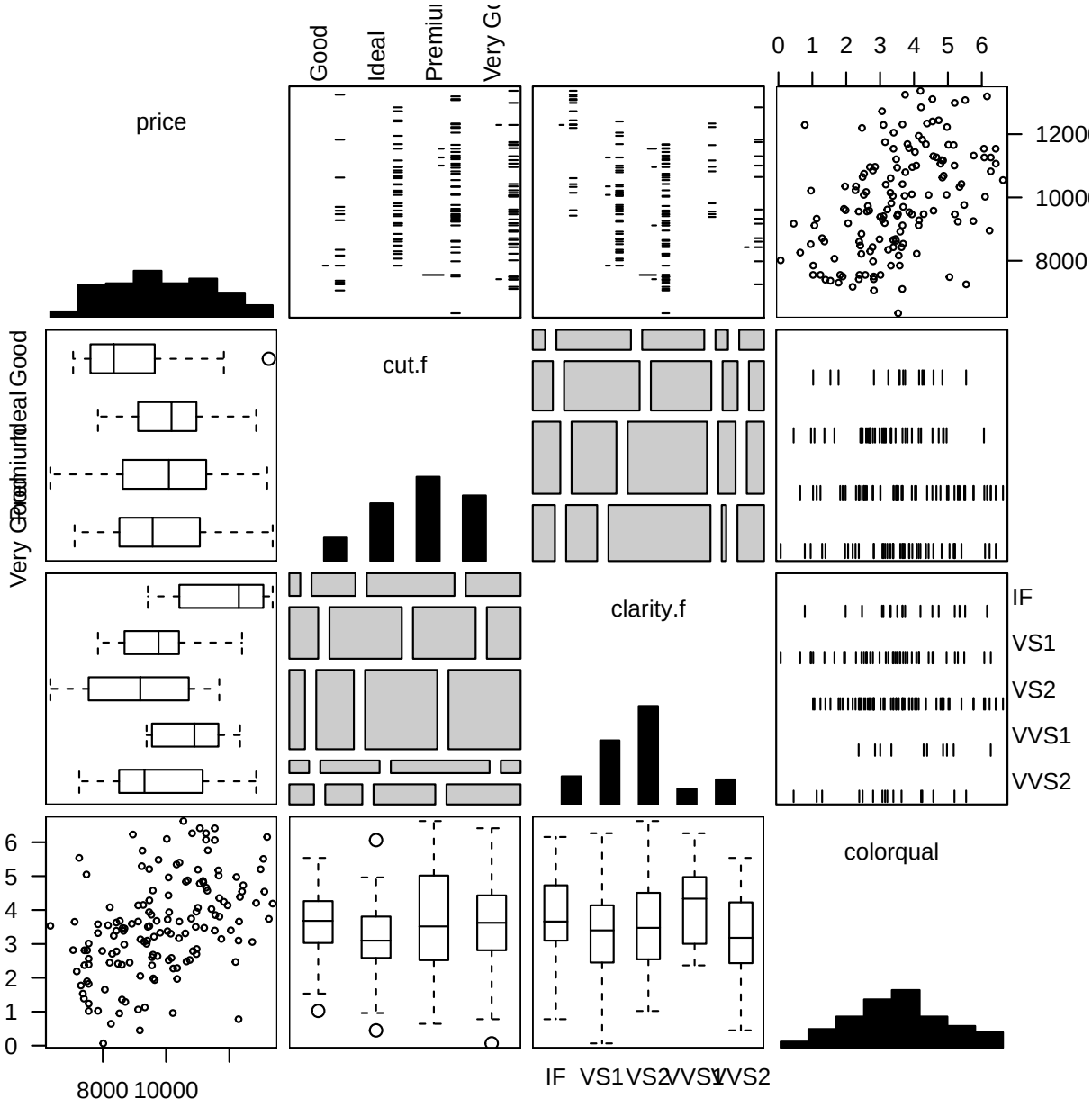


Figure 3: A generalize pairs plot, in case you find it helpful.

After excluding cut, she obtains the following model. Various residual plots for `lm.full` and `lm.nocut` appear in Figure 4.

```
> lm.nocut <- lm(price ~ clarity.f + colorqual, data=x)
> summary(lm.nocut)

Call:
lm(formula = price ~ clarity.f + colorqual, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-3680.0  -966.6   -22.9   908.5  2571.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    9998.40     406.58  24.591  < 2e-16 ***
clarity.fVS1  -1820.54     356.41  -5.108 1.03e-06 ***
clarity.fVS2  -2526.54     335.16  -7.538 5.14e-12 ***
clarity.fVVS1 -1312.84     494.72  -2.654  0.00887 **
clarity.fVVS2 -1805.28     432.79  -4.171 5.25e-05 ***
colorqual       495.12      73.74   6.714 4.20e-10 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 1253 on 142 degrees of freedom
Multiple R-squared:  0.4417,        Adjusted R-squared:  0.422
F-statistic: 22.47 on 5 and 142 DF,  p-value: < 2.2e-16

> anova(lm.nocut, lm.full)

Analysis of Variance Table

Model 1: price ~ clarity.f + colorqual
Model 2: price ~ clarity.f + colorqual + cut.f
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    142 222783212
2    139 206548944  3  16234268 3.6417 0.0144 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

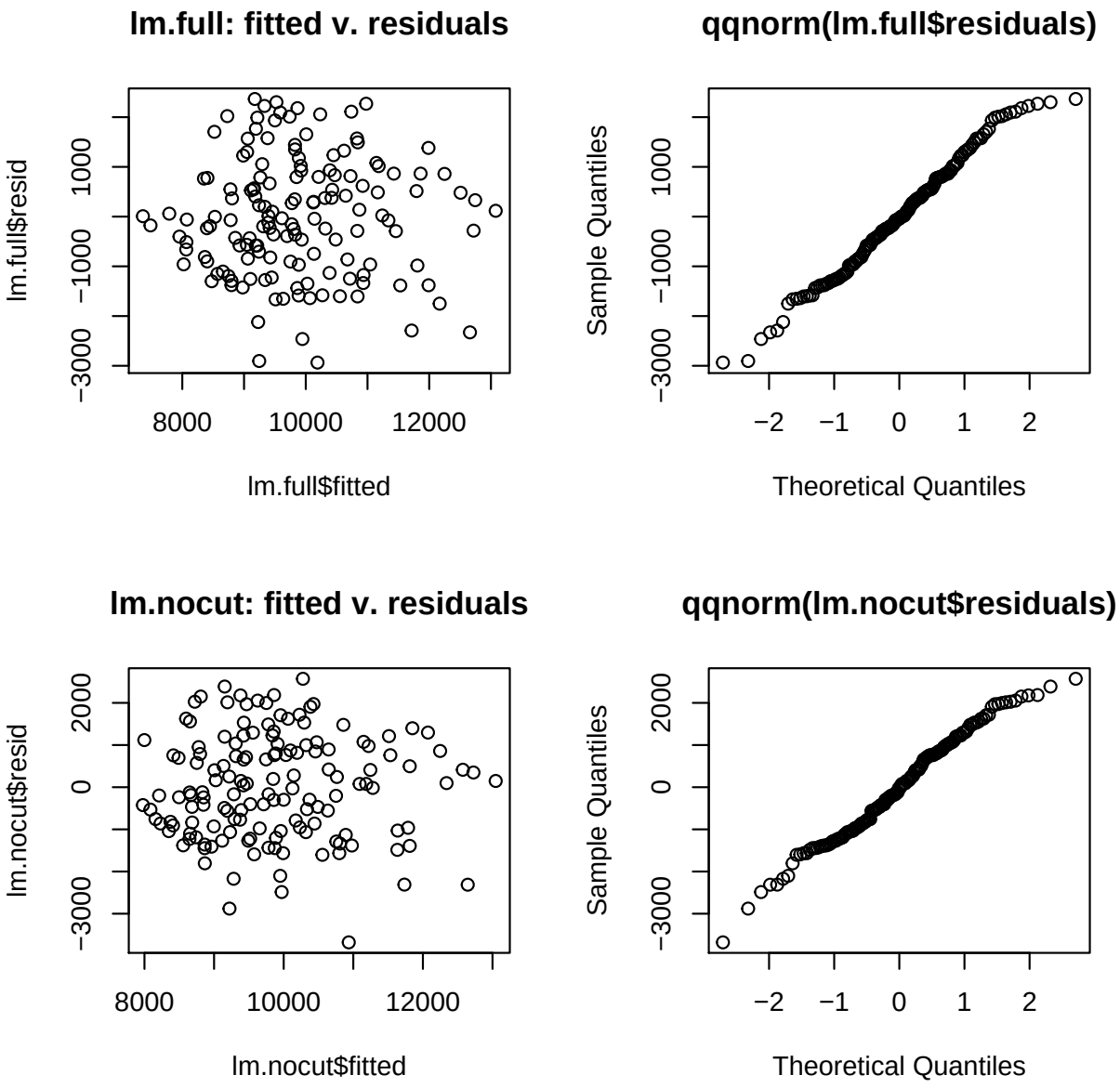


Figure 4: Residual plots for `lm.full` and `lm.nocut`.

## FINAL QUESTIONS ABOUT DIAMONDS!

**Question 7A:** Which do you prefer, `lm.full` or `lm.nocut`? Explain, with specific reference to at least one numerical quantity (a test statistic, p-value, etc...).

**Question 7B:** Do you see anything in Figure 4 that would lead you to prefer either `lm.full` or `lm.nocut`? Explain.

**Question 7C:** What is the predicted price of a diamond of "Ideal" cut, with `colorqual` equal to 2.0, and of clarity "IF", using model `lm.full`? Show the algebra.

**Question 7D:** Explain to the diamond merchant (who took Intro Stats many years ago and is rusty) the meaning of "Multiple R-squared" equal to 0.4417 in model `lm.nocut`. This should be on the order of magnitude of a tweet; a paragraph is unnecessary.