Final exam practice

Joshua Loftus

Study recommendations

- Read the lecture notes posted on the course page. You can mostly skip over `R` code, since there will be no coding on the exam. You should look at the output from the code, like test and confidence interval output, or plots. These notes should be sufficient to answer any exam problems, but you can also review the textbook references linked on the course page.
- Review homework solutions.
- Any topics covered in lectures may appear on the exam. This includes, for example:
  - Concepts related to observational studies
  - Probability and random variables
  - Sampling distribution of the mean
  - Parameter estimation and unbiasedness
  - Mean-squared error and bias-variance tradeoff
  - Law of large numbers and central limit theorem
  - Confidence intervals and hypothesis tests for a mean, for a difference in means, and a hypothesis test for independence
  - Relationship between sample size, confidence level, and width of confidence interval
  - Null and alternative hypotheses
  - Relationship between intervals and tests
  - Significance level and confidence level
  - Type 1 and type 2 errors
  - Definition, interpretation, and calculation of p-values
  - Use and misuse of p-values
  - Relationship between effect size, sample size, and power
  - Basic idea of how to calculate a large enough sample size
  - Covariance, correlation, and linear relationships
  - Regression lines and relationship with correlation
  - Interpretation and estimation of regression model coefficients
  - Interpretation of coefficients in multiple regression models
  - Interpretation of coefficients for categorical predictors
  - Reading summaries: estimates, p-values, adjusted R-squared, confidence intervals
  - Using diagnostic plots to detect problems with linear models
  - Interpretation of F-test for nested models
  - (The following topics were only covered at a high level, so they may only be tested for basic concepts and not technical details)
  - Model selection concepts including overfitting, penalizing complexity, using training and test data, and the relationships between complexity, bias, variance, and accuracy on the training vs test sets
  - Purpose of logistic regression (predicting probability of a categorical outcome being 0 or 1)
  - Collinearity, Simpson's paradox, association vs causation, measurement

Question 1

We will fit a linear model using the `diamonds` dataset. There are two continuous variables, `price` and `carat` (the weight of the diamond), and three categorical variables (`cut`, `color`, `clarity`). The first few rows of data are shown below.

```
## # A tibble: 6 x 5
##   carat cut     color clarity price
##   <dbl> <fct>   <fct> <fct>   <int>
## 1 0.230 Premium Good  Bad       326
## 2 0.210 Premium Good  Bad       326
## 3 0.230 OK      Good  Good      327
## 4 0.290 Premium Bad   Good      334
## 5 0.310 OK      Bad   Bad       335
## 6 0.240 OK      Bad   Good      336
```

There are 53940 rows in this dataset, each one is an observation corresponding to an individual diamond. Before analyzing the data, we first split it into two random subsets, a **training set** with 5394 observations and a **test set** with 29976 observations.

Part (a)

```
##
## Call:
## lm(formula = log(price) ~ carat + cut, data = dtrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1155 -0.2558  0.0371  0.2612  1.4082
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.20247    0.01350 459.585  < 2e-16 ***
## carat        1.94227    0.01146 169.440  < 2e-16 ***
## cutPremium   0.04834    0.01161   4.164 3.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4035 on 5391 degrees of freedom
## Multiple R-squared:  0.842,  Adjusted R-squared:  0.842
## F-statistic: 1.437e+04 on 2 and 5391 DF,  p-value: < 2.2e-16
```

- What outcome variable is this model predicting?

- What are the predictor variables?

- What is the interpretation of the coefficient for `carat`? Does the fact that this coefficient is positive surprise you, or is it what you would expect?

- What are the interpretations for the coefficients of `(Intercept)` and `cutPremium`? Which one is larger, and does that surprise you or is it expected?

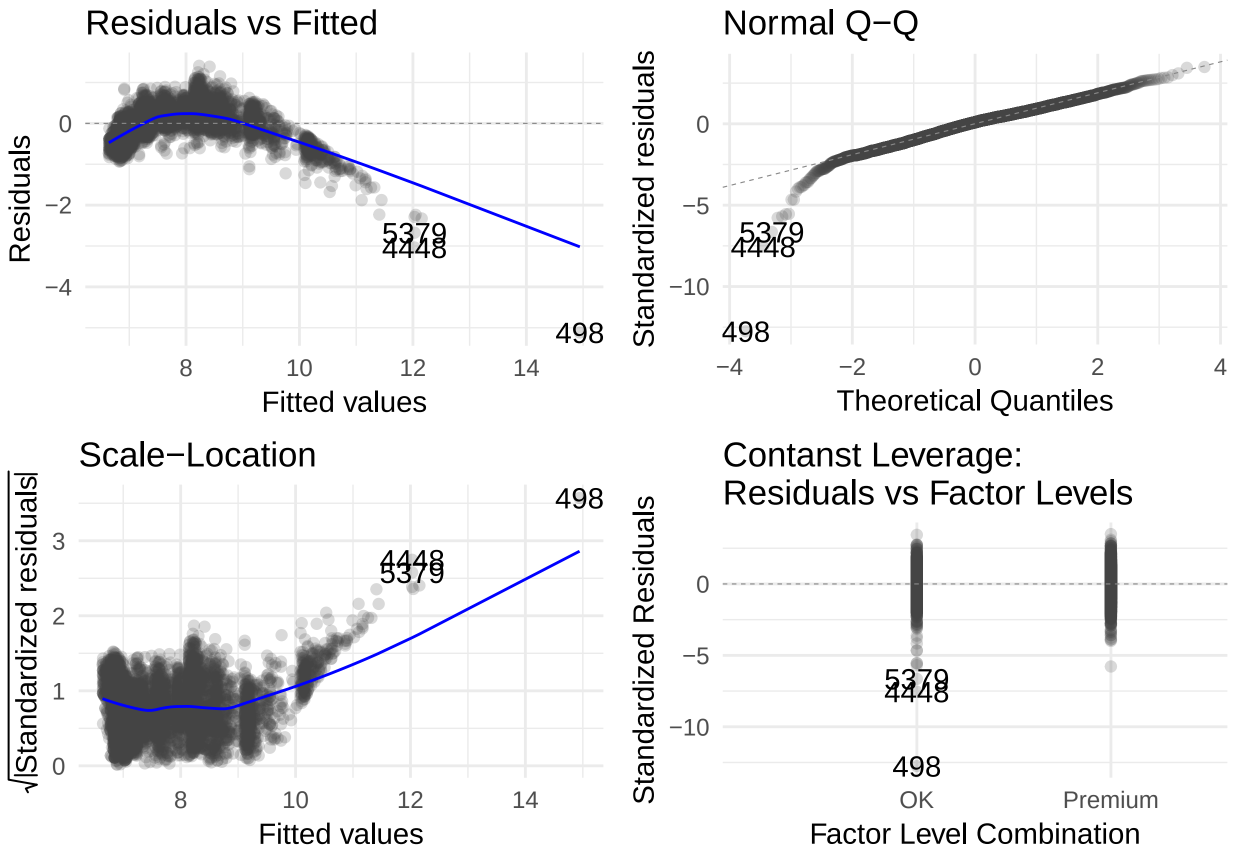
- Why are all the p-values extremely small even though the coefficients are not very large?

- What is the null hypothesis for the p-value in the first row of the summary (the row for the `carat` variable)?

- What is the null hypothesis for the p-value in the last line of the summary (the line with an F-statistic)?

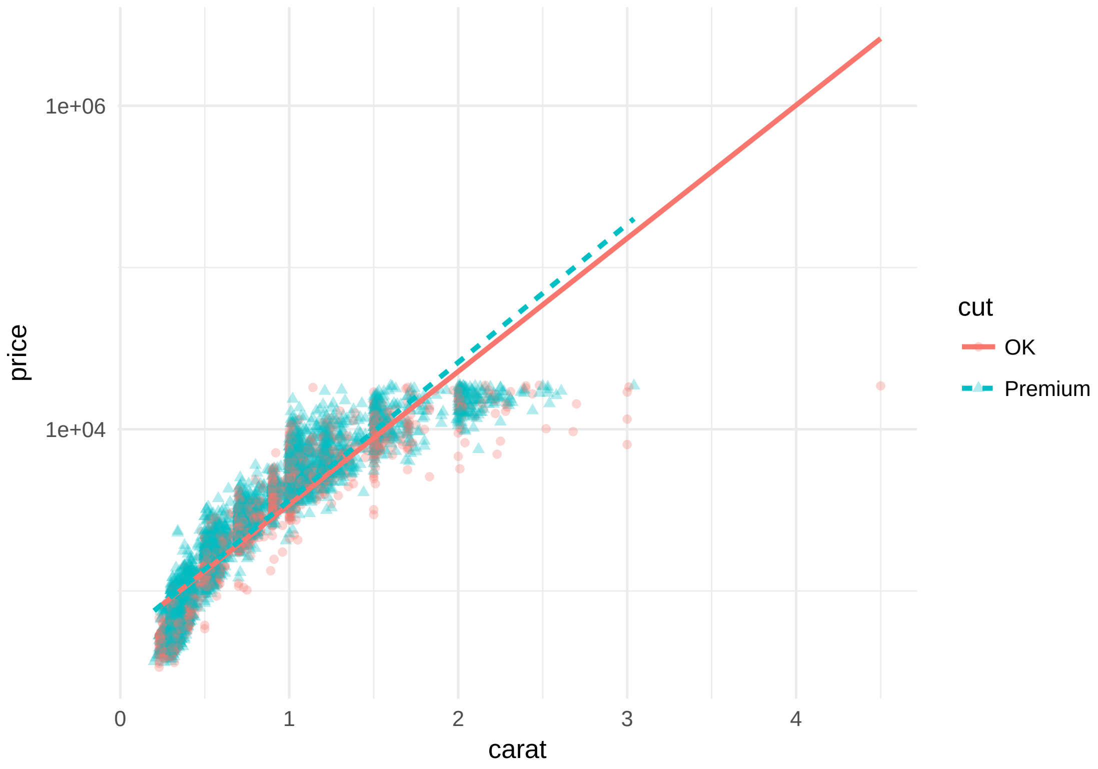
- How would you interpret the adjusted R-squared for this model?

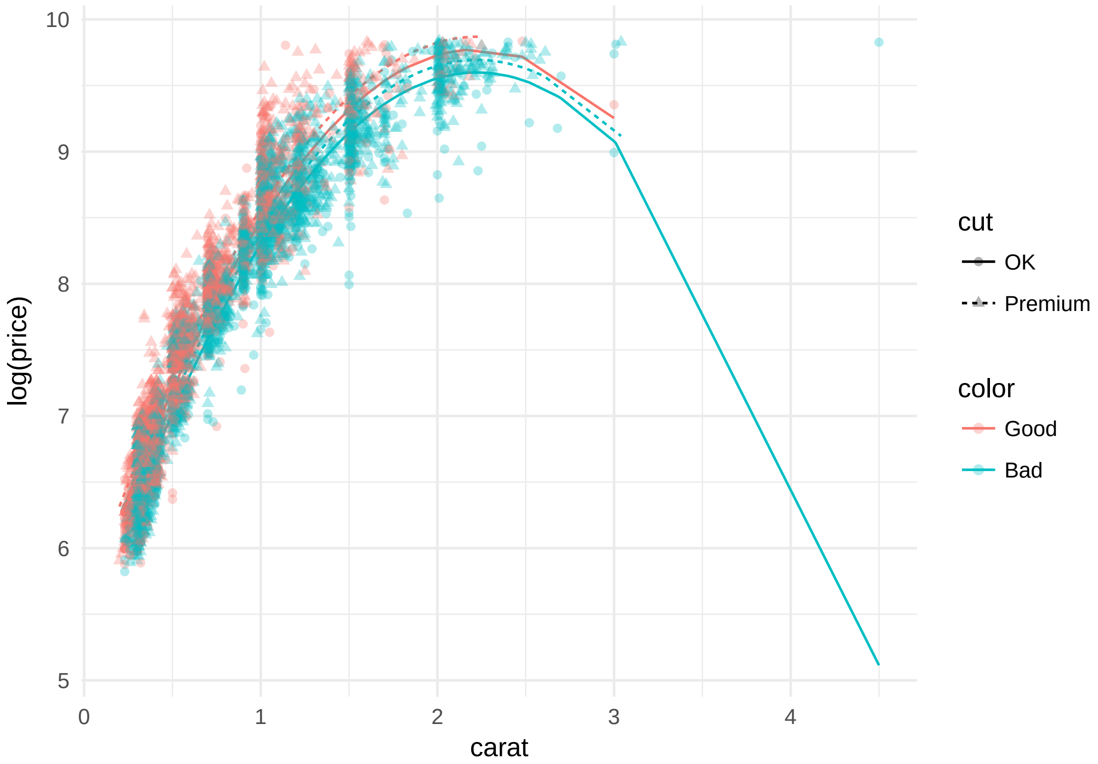
Part (b)



- What problems can you identify with the model based on the above diagnostic plots? Describe each problem with one sentence.

- If you could do one thing to try to fix these problems, what would it be? (Hint: see the **first plot** below)


price
1e+06
1e+04
0
1
2
3
4
carat
cut
OK
Premium


log(price)
10
9
8
7
6
5
0
1
2
3
4
carat
cut
OK
Premium
color
Good
Bad

Part (c)

- The **second plot** above shows four lines which are predictions given by a new model, Model 2. What predictor variables are in this new model? (Hint: there are more than one new variables included)

- The table below shows adjusted R-squared and residual sums of squares (test error) on the test set for three models in increasing order of model complexity. Why does the most complex model, Model 3, have the best adjusted R-squared but also a larger test error than the other models?

```
##              Model 1  Model 2   Model 3
## AdjRsquared    0.842    0.935     0.965
## TestRSS     5386.943 2248.667 67031.673
```

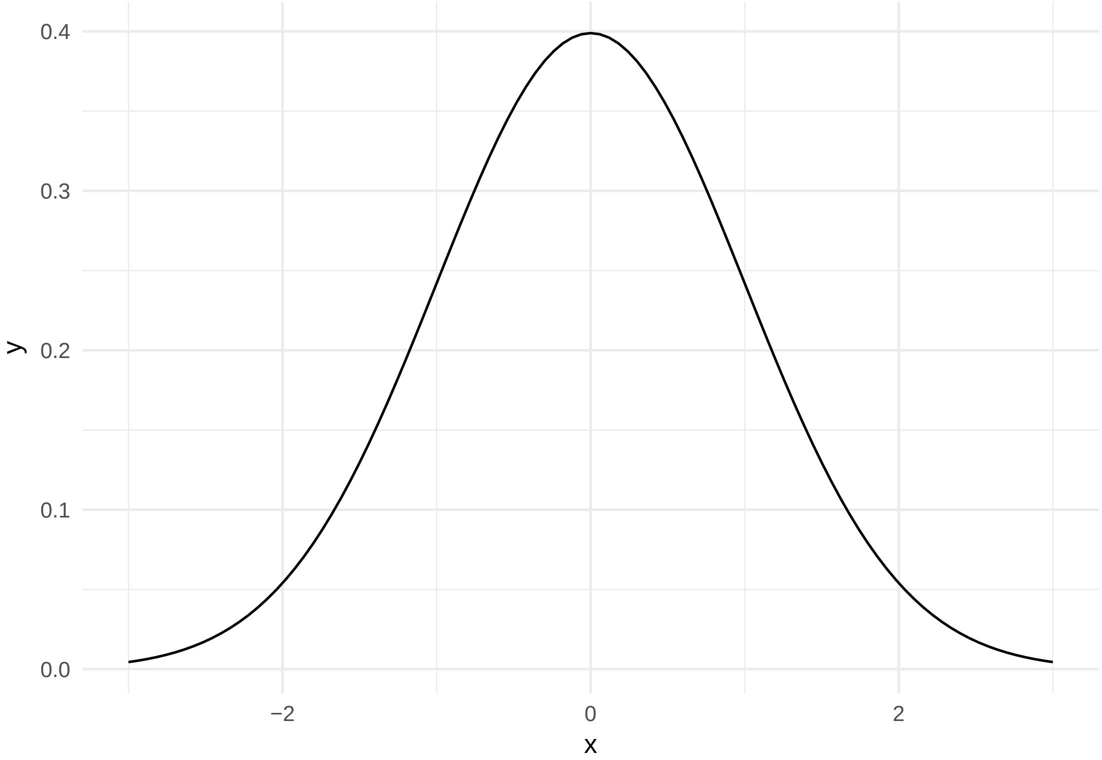
- Sketch a graph below to show the relationship between model complexity, bias-squared, and variance. Label your horizontal axis and each line in the graph. Pick three values on the horizontal axis, one for each model above, and label them Model 1, Model 2, and Model 3.

Question 2

In this question we are interested hypothesis tests and confidence intervals for a coefficient β in a linear model.

Part (a)

Suppose we test $H_0 : \beta = 0$ against the one-sided alternative $H_1 : \beta > 0$, and the p-value we obtain is 0.7. In the graph below, shade the region that (approximately) corresponds to this p-value.



Part (b)

Suppose we switch to the opposite one-sided alternative $H_1 : \beta < 0$. Would this alternative give us a different p-value than the previous one? If not, why not? If yes, what is the new value?

Part (c)

Suppose that β is included in a model with several other coefficients, β_1 and β_2, and we use an F-test to test the hypothesis $H_0 : \beta = \beta_1 = \beta_2 = 0$. If we reject this null hypothesis, does it mean that β must be nonzero?

Part (d)

Suppose that $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 < 0$. Does this mean the correlation between the outcome variable y and the predictor x_1 is positive, and $\text{cor}(y, x_2) < 0$? Explain.

Part (e)

Suppose observations in the dataset correspond to small geographic regions like neighborhoods, y is the rate of emergency room visits from that neighborhood due to asthma, and x_1 measures the concentration of air pollution. True or false, and explain: β_1 is the increase in rate of ER visits due to asthma caused by pollution, holding other predictors constant.