Midterm 1 Study Guide

ECON 4753 — University of Arkansas

Prof. Kyle Butts

Overview of Topics:

1. Statistics Review
   - Understand the intuition of a sample distribution, i.e. repeated sampling ($*$)
   - Say you have some *statistic* that has a normal sample distribution, e.g. $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. You should know these three inference procedures:
     - Form a 95% confidence interval
     - Perform a hypothesis test that $H_0 : \mu = m$ for some number m (e.g. $H_0 : \mu = 0$)
     - Construct a rejection region (i.e. outside of the confidence interval centered on the null hypothesis)

2. Introduction to Forecasting
   - High-level understanding of forecasting, $y = f(X) + u$
   - Two goals of forecasting: prediction and inference
   - What model "flexibility" means and the risk of overfitting
   - Mean-squared prediction error
   - Differences between training data and testing data

3. Bivariate Regression
   - Understand how to interpret correlation coefficient
   - The regression coefficients: $\hat{\beta}_1 = \frac{\text{cov}(X,y)}{\text{var}(X)}$ and $\hat{\beta}_0 = \bar{y} - \bar{X}\hat{\beta}_1$
   - Given a regression coefficient, how do we interpret the marginal effect
     - A one unit change in X is associated with a $\hat{\beta}_1$ change in y
     - DO NOT use causal langauge here
   - Understand regressing y on an indicator variable
     - How do you interpret the intercept and the coefficient on the indicator variable

(difference-in-means)

- Regression y on a set of indicator variables for each value of a discrete variable
  - Know what the 'omitted' category means and know the coefficients are difference-in-means
- $\log$ transformations
  - Percent change in $\log$-transformed variables

4. Multiple Regression
   - Understand the motivations for multiple regression
   - Understand the intuition of "all else equal"
   - Interpret a polynomial regression
   - Interpret interactions of binary variables
     - E.g. gender, college degree, and their interaction

Study Questions

Statistics Review

1. Say a sample mean of 64 individuals was $\bar{X} = 24$ and the sample variance is 9. Construct a 95% confidence interval for the sample mean.
2. Say a sample mean of 64 individuals was $\bar{X} = 24$ and the sample variance is 9. Can you say the sample mean is statistically significantly different from $\mu = 20$ at the $\alpha = 0.05$ level.
3. Say a sample mean of 36 individuals was $\bar{X} = 3$ and the sample variance is 4. Is this estimate significantlly different from $\mu = 0$?
4. Say I calculate a rejection region for a $\alpha = 0.05$ significance level to be $\bar{X} \leq 16$ or $\bar{X} \geq 24$. In words, what does this rejection region represent?

Introduction to Forecasting

1. In words, how do the goals of forecasting, 'prediction' and 'inference', differ?
2. Say you are using a model of $y = f(X) + u$ for a single variable X. Give two reasons why you might want to use a linear model for $f(X)$.
3. The MSPE is given by $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(X))^2$. Why might it be a bad idea to pick a model for f based on the in-sample MSPE?
4. Calculate the mean-squared prediction error for the following sample:

y_i	$\hat{y}_i$
3.7	4.20
4.1	4.18
5.6	5.48
2.9	3.29
8.8	8.81

5. In words, describe the bias-variance trade-off when describing the flexiblity of $f(X)$

Bivariate Regression

1. Say the $\text{cov}(X, y) = 4.2$, $\text{var}(X) = 2.1$, $\bar{X} = 10$, and $\bar{y} = 4$. What are the ordinary least squares estimate for this data?
2. Here we have a survey of workers in 1995. We regress annual earnings on a worker's age

```
OLS estimation, Dep. Var.: annual_earnings
Observations: 15,992
             Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)     31369      238.95    131.28  < 2.2e-16 ***
age            264.75       14.82    17.858  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   i. What is the predicted annual earnings for a 30 year old worker?
   ii. In words, what is the marginal effect of `age`? Is this relationship statistically significant?
   iii. Can you reject the null that the marginal effects of age is \$240 per year at a $\alpha = 0.05$ level of significance?
   iv. Why might we want to make our linear model into a more flexible model?
3. Consider the following regression using the `mtcars` dataset. This is a cross-sectional dataset set of 32 cars with info on their miles per galleon (`mpg`) and the number of cylinders in their engine (`cyl`). A car can have either 4, 6, or 8 cylinders.

```
OLS estimation, Dep. Var.: mpg
Observations: 32
             Estimate  Std. Error   t value    Pr(>|t|)
(Intercept)  26.66364    0.971801  27.43735   < 2.2e-16 ***
cyl::6       -6.92078    1.558348  -4.44110  1.1947e-04 ***
cyl::8      -11.56364    1.298623  -8.90453  8.5682e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   i. Interpret in words the coefficient on `cyl::6`.

ii. What is the mean mpg for 8-cylinder cars?

Multiple Regression

1. Consider the following regression using the `mtcars` dataset. This is a cross-sectional dataset set of 32 cars with info on their miles per galleon (`mpg`), their horsepower (`hp`), their weight (`wt`), and the number of cylinders in their engine (`cyl`)

```
OLS estimation, Dep. Var.: mpg
Observations: 32
             Estimate  Std. Error   t value    Pr(>|t|)
(Intercept) 35.845995    2.041019  17.56279  2.6703e-16 ***
hp          -0.023120    0.011952  -1.93436  6.3613e-02 .
wt          -3.181404    0.719601  -4.42107  1.4418e-04 ***
cyl::6      -3.359025    1.401670  -2.39645  2.3747e-02 *
cyl::8      -3.185884    2.170475  -1.46783  1.5370e-01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   i. What is the estimated marginal effect of an additional horsepower on the miles per gallon?
   ii. Holding fixed horse-power and the number of cylinders, do heavier cars have higher or lower mpg? Is this relationship statistically significant?
   iii. Form a 95% confidence around the slope coefficient on `wt`

2. Here we have a survey of workers in 1995. We regress annual earnings on an indicator for being a Black workers, an indicator for having a high-school degree, and an interaction between the two.

```
OLS estimation, Dep. Var.: annual_earnings
Observations: 15,992
                Estimate  Std. Error   t value    Pr(>|t|)
(Intercept)     32817.225     146.402  224.1583   < 2.2e-16 ***
black           -2152.322     445.900  -4.82691  1.3995e-06 ***
```

```
hs_degree        3153.177     173.126  18.21323   < 2.2e-16 ***
black:hs_degree  -783.909     585.840  -1.33809     0.18088
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

i. What is the "omitted group" in this?

ii. How would you interpret the coefficient on 'black $\times$ hs degree'?

iii. Do the benefits of having a high-school degree signficiantly differ for Black and non-Black workers?

Example Exam

Here is an example of what an exam could look like:

1. Say you collect data on a sample of $n = 36$ films produced by A24. For each film, you record X_i to be the rotten tomatoes review score. You calculate the sample average score to be 84% and a sample variance of 9%.
   - What is the sample distribution of the sample mean $\bar{X}$?
   - Write a 95% confidence for your sample mean estimate (the critical value of the middle 95% is ± 1.96). In words, describe your confidence interval.
   - Conduct a hypothesis test that the true average review is $H_0 : \mu = 85.5\%$ at the $\alpha = 0.05$ level of significance.
2. Continuing our example, say you want to predict a model for how a film may be reviewed. You collect data on the genre of the film (say the categories are action, drama, and comedy). You regress the film's review on an intercept an indicator for action and an indicator for drama.

```
                  Estimate Std. Error  t value   Pr(>|t|)
(Intercept)          76.66       4.97   15.424  < 2.2e-16 ***
genre::"action"       6.92       3.75    1.845      0.064 .
genre::"drama"        9.56       3.30    2.897      0.004 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   - What is the average rating for drama movies?
   - What is the "omitted group" in this regression?
   - Is it true that drama films have statistically significantly higher reviews than comedy films? How do you know?
3. Say you were hired by A24 to be a data scientist. They give you data on all of their films and some extra varaibles. They want you to continue to predict what makes a movie review well
   - They tell you they prefer you to do 'inference' instead of 'prediction'. Why might that be?
   - Now say you run a model and you want it to be 'flexible' in predicting reviews. Why might it be useful for you to set some movies aside to be your 'test dataset'?

- Say you have two key variables you want to include: the budget for the film and how many award-winning actors the film features. Why might you want to run a multivariable regression for this problem?