Practice Final

Name: Solutions

You may use a calculator and three 8.5" by 11" sheets of notes (front and back), which you **will** turn in with your exam. This means you have to go to an **open book** room to take the test. However, you may **not** use the text book. You will have to bring your own calculator.

Please show all your work, including all calculations, and explain your answers. Whenever needed, please round numbers (including intermediate calculations) to the nearest 0.001.

Cell phones and any other electronic devices (aside from your calculator) are not permitted. No interaction of any sort is allowed with your classmates.

I have tried to be very clear in my statements of all questions on this exam. If there are any questions where it is not clear what I am asking, please write down your best guess at what I am asking and answer that.

I Conceptual Questions

Please answer the following in no more than 1-2 sentences each.

1. (4 points) Comment on the following statement attributed to statistician George Box:

   All models are wrong but some are useful.

   How does this statement relate to what we have learned about linear regression?

The linear model makes several assumptions: No outliers, a linear relationship, independence, normally distributed errors, and equal variance of the errors. In a real data set, it's rare that all of these assumptions are exactly satisfied. However, these assumptions are often approximately correct. If the assumptions are "close enough" to being true the model can still be a useful way to learn about the relationship between the explanatory and response variables, even if the model is "wrong".

2. (4 points) What does it mean for data to be **paired**?

If we are doing inference for a difference in means between two populations the data are paired if there is some sort of natural connection between the units or people we have sampled from these two populations. For example if we measure some characteristic (like blood pressure) on the same individuals both before and after they have been exposed to some treatment (like taking a medicine) those observations are paired because there is a natural connection or dependence between the before and after values measured on the same person.

3. (4 points) What is a sampling distribution?

A sampling distribution is the distribution of values for a certain statistic across all possible samples of a certain size from the population.

4. (6 points) Suppose I conduct a hypothesis test about the average amount of tea in a Twinings tea bag. The null hypothesis is that the population mean amount of tea is (less than or) equal to 2.5 grams, and the alternative hypothesis is that the population mean amount of tea is larger than 2.5 grams. In this context, what would a Type I error be? If I change the significance level of the test, α, from 0.05 to 0.01, how does that affect the probability of making a Type I error in this test?

A Type I error is the error of rejecting the null hypothesis incorrectly when the null hypothesis is actually true. In this example, we would make a type I error if the mean amount of tea was actually 2.5 grams, but we rejected the null hypothesis and concluded that the mean amount of tea was greater than 2.5 grams. If the null hypothesis is true the probability of making a Type I error is equal to the significance level α; in that case reducing α from 0.05 to 0.01 results in a lower probability of type I error.

5. (3 points) A survey of 500 households concluded that 82% of the population uses coupons at the grocery store. Describe what is meant by the poll having a margin of error of 3%.

The margin of error describes our uncertainty in our estimate of the proportion of the population who use coupons at the grocery store. The margin of error can be used to construct a confidence interval for the population proportion (typically a 95% confidence interval). In this case, based on the survey results we are 95% confident that between 79% and 85% of the population use coupons.

II Applied Problems

1. (18 points) We are interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

   (a) (2 points) Describe the population parameter of interest.

   The population parameter is the proportion of all 4500 students in the graduating class who found jobs within 1 year of graduating.

   (b) (4 points) Check if the conditions for constructing a confidence interval and conducting a hypothesis test based on these data are met.

   - 2 outcomes: find a job or did not
   - same prob. of success for each randomly selected person in our sample,
   - independence:
     - randomization used in selecting graduates to include in the sample
     - sample size (400) is < 10% of population size (4500)
   - we have at least 10 successes and at least 10 failures in our data set;

   (c) (4 points) What is a 95% confidence interval for the proportion of graduates who found a job within one year of completing their undergraduate degree at this university? Interpret the interval **in the context of this problem.**

   $\hat{p} = \frac{348}{400} = 0.87$ $\quad SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.87 \cdot 0.13}{400}} = 0.017$

   $\hat{p} \pm 2 \cdot 0.017 \rightarrow [0.87 - 0.034,\ 0.87 + 0.034)$
   $\rightarrow [0.836,\ 0.904)$

   We are 95% confident that the proportion of all graduates who find a job within 1 year is between 0.836 and 0.904.

   (Note that I did not specifically ask you to explain what you mean by 95% confident. If I had asked you that you'd say: If we took many different samples and computed a separate 95% C.I. for the population proportion based on the data in each sample, about 95% of these confidence intervals would contain the population proportion.)

(d) (4 points) According to the National Center for Education Statistics, the proportion of all 20-to-24 year olds with college degrees who were employed as of 2015 is 0.89. Write down a statement of the null and alternative hypotheses for a test that the employment rate for the current graduating class is different from the national average among 20 to 24 year olds.

$H_0: p = 0.89$

$H_A: p \neq 0.89$

(e) (4 points) I used R to compute a p-value for this test, and I got a p-value of 0.2. **In the context of this problem**, what is your conclusion? Use a significance level of $\alpha = 0.05$.

Since the p-value of 0.2 is greater than $\alpha = 0.05$ we fail to reject the null hypothesis. The data do not offer enough evidence to conclude that the proportion of graduates in this class who found a job within 1 year is different from the proportion of all 20-to-24 year olds with college degrees who were employed as of 2015.

2. (16 points) Researchers interested in lead exposure due to car exhaust sampled the blood of 52 police officers subjected to constant inhalation of automobile exhaust fumes while working traffic enforcement in a primarily urban environment. The blood samples of these officers had an average lead concentration of 124.32 micrograms/liter and a SD of 37.74 micrograms/liter; a previous study of a large number of individuals (not all police officers) from a nearby suburb, with no history of exposure, found an average blood level concentration of 35 micrograms/liter.

   (a) (3 points) Write down the hypotheses that would be appropriate for testing if the police officers appear to have been exposed to a higher concentration of lead than their neighbors in the suburbs. You may treat the value of 35 micrograms/liter from the suburban study as a fixed, known constant (i.e., we are not structuring this as a test to compare the means of two groups, but rather as a test about the mean value for the group of urban police officers).

   $H_0: \mu = 35$

   $H_A: \mu \neq 35$ (or $H_A: \mu > 35$ if you prefer to do a 1-sided test).

   μ is the mean blood concentration of lead in the population of all urban police officers.

   (b) (4 points) Explicitly state and check all conditions necessary for inference on these data. If you don't have enough information, say what you would want to know. If you would want to look at any plots, describe what plot(s) you would want to make and what you'd be looking for.

- independence: we don't have enough information to assess this, we would need to know that the police officers were selected randomly, and that the # of police officers in our sample was less than 10% of the # of police officers in the sample.
- normal distribution: I'd need to check a histogram or density plot of the lead concentrations in the blood samples to assess this condition. I would want to see a distribution that was unimodal and approximately symmetric.
- sample size: We have 52 blood samples. That's probably enough to ensure that the sampling distribution of the mean will be approximately normal, if the distribution of values is symmetric and unimodal enough that reporting the mean as a summary of the center makes sense in the first place.

(c) (4 points) Calculate a p-value for the test that the downtown police officers have a higher lead exposure than the group in the previous study. **For full credit, you must show all of your work!** Points are assigned to correct set-up. In doing this calculation, you may use the following facts:

- ~~If $T \sim t_{51}$, then $P(T > 23.754) < 10^{-16}$~~
- If $T \sim t_{51}$, then $P(T > 17.067) < 10^{-16}$
- ~~If $T \sim t_{51}$, then $P(T > 3.294) = 0.001$~~
- If $T \sim t_{51}$, then $P(T > 2.367) = 0.011$

$\frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{124.32 - 35}{37.74/\sqrt{52}} = 17.067.$

If the null hypothesis is true, this follows a t_{n-1} distribution.

If we use a 1-sided alternative like $H_A: \mu > 35$, the p-value is the probability above something less than 10^{-16}. If we use a 2-sided alternative like $H_A: \mu \neq 35$, the p-value is twice that value.

Either way, the p-value is very small.

(d) (3 points) Interpret your results in context.

Since the p-value is $< \alpha = 0.05$, we reject the null hypothesis. The data offer enough evidence to conclude that the urban police officers do not have a mean blood lead concentration of 35 micrograms/liter.

(e) (2 points) Suppose that you rejected the null hypothesis. Would this prove that there was a causal relationship between exposure to car exhaust and increased concentrations of lead in the blood?

No, we cannot assert a causal relationship based on the analysis of these observational data. There could be lurking variables or other factors that explain the difference— for example, maybe the police officers in the current study handle lead bullets more than the non-police officers in the previous study.

3. (19 points total) An experiment conducted by the MythBusters, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously influenced into yawning if another person near them yawns. They recruited 50 people and randomly assigned them to two groups: 68 to a group where a person near them yawned (treatment) and 16 to a group where there wasn't a person yawning near them (control). The following table shows the results of this experiment.

		Group		
		Treatment	Control	Total
Result	Yawn	10	4	14
	Not Yawn	24	12	36
	Total	34	16	50

(a) (2 points) The MythBusters did not conduct a formal statistical analysis of these experimental results. Reproduce their analysis here by calculating the proportion of subjects in the treatment group who yawned, and the proportion of subjects in the control group who yawned.

In The treatment group $\frac{10}{34} = 0.294$ is the proportion of subjects who yawned.
In the control group, the proportion of subjects who yawned is $\frac{4}{16} = 0.25$.

(b) (2 points) Which group had a higher proportion of subjects who yawned? Could this result have occurred by chance even if on average the two groups had the same chances of yawning?

~~Maybe~~ A higher proportion of subjects in the treatment group yawned but this could have occurred by chance even if the two groups had the ~~the~~ same chances of yawning.

(c) (3 points) Check the necessary assumptions for obtaining confidence intervals and conducting hypothesis tests with these data.

Independence:
- subjects were randomly selected & randomly assigned to treatment groups.
- sample sizes are less than 10% of the population size (population is all humans)

2 outcomes: either subject yawns or not

Same probability of success for each trial within each group: seems reasonable.

Independence across the two groups: subjects were kept separate from each other, so they would not influence one another.

You may use the following R output in answering the questions on the following page (note that the only difference between the following commands is the specification of the alternative hypothesis).

```
prop.test(x = c(30, 8), n = c(68, 32), conf.level = 0.99, alternative = "two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(30, 8) out of c(68, 32)
## X-squared = 2.6, df = 1, p-value = 0.1
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.08266  0.46502
## sample estimates:
## prop 1 prop 2
## 0.4412 0.2500
```

```
prop.test(x = c(30, 8), n = c(68, 32), conf.level = 0.99, alternative = "greater")
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(30, 8) out of c(68, 32)
## X-squared = 2.6, df = 1, p-value = 0.05
## alternative hypothesis: greater
## 99 percent confidence interval:
##  -0.05837  1.00000
## sample estimates:
## prop 1 prop 2
## 0.4412 0.2500
```

```
prop.test(x = c(30, 8), n = c(68, 32), conf.level = 0.99, alternative = "less")
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(30, 8) out of c(68, 32)
## X-squared = 2.6, df = 1, p-value = 0.9
## alternative hypothesis: less
## 99 percent confidence interval:
##  -1.0000  0.4407
## sample estimates:
## prop 1 prop 2
## 0.4412 0.2500
```

(d) (8 points) Regardless of your answer to part (c), let's go ahead with inference. State and interpret a 99% confidence interval that is relevant to answering the question of whether a person can be subconsciously influenced into yawning if another person near them yawns. As part of your answer, discuss what it means to be "99% confident".

We are 99% confident that the difference in chances that someone will yawn if someone near them yawns vs. if no one near them has yawned is between −0.083 and 0.465. If we were to collect data from many different samples and use the data from each sample to construct a separate 99% confidence interval, about 99% of those confidence intervals would contain the true difference in probability that someone will yawn if someone near them has yawned vs. if no one near them has yawned, in the population.

(e) (8 points) Suppose you'd like conduct a hypothesis test to see if yawning is contagious. Carry out a suitable hypothesis test at an $\alpha = 0.05$ significance level. Clearly state your null and alternative hypotheses, defining all population parameters that are involved. State your conclusions in context. Do you agree with the MythBusters' conclusion that these data provide strong evidence that people are more likely to yawn if someone near they yawns?

$H_0: p_1 = p_2$ p_1 = probability someone will yawn if someone near them yawned recently

$H_A: p_1 > p_2$ p_2 = " " " " " " no one yawned near them recently.

The p-value for this test is 0.05, which is right on the border between rejecting and failing to reject the null hypothesis! (if using $\alpha = 0.05$). Technically we fail to reject H_0 in this case. The data do not provide enough evidence to conclude that people are more likely to yawn if someone near them yawned recently.

actually, if you compute the p-value to 3 decimals, it's 0.053 so we definitely fail to reject.

4. (30 points) What is the relationship between a movie's budget and its earnings? Let's look at data for 35 recent Action and Adventure movies.

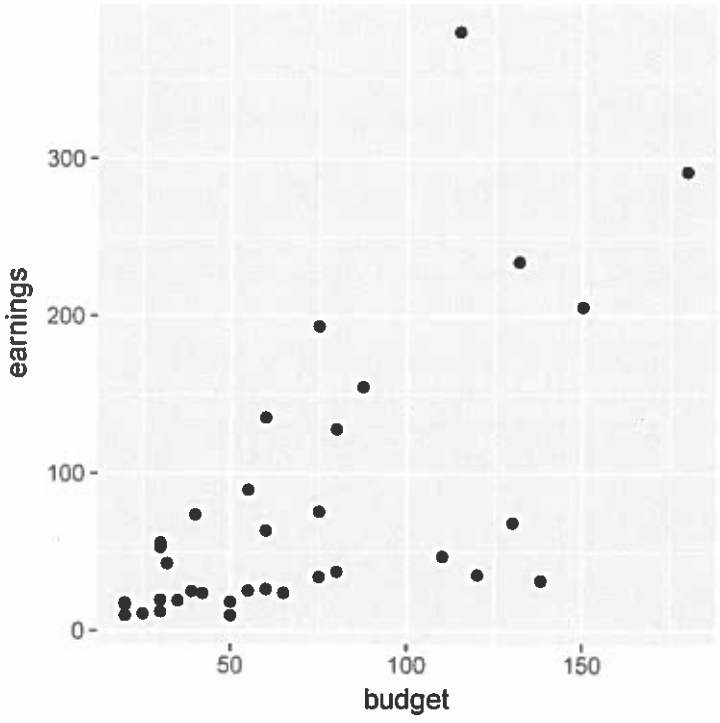
```
## A first look
head(movies)
```

```
##                     title budget earnings     genre
## 1                 Elektra     65    24.41    Action
## 2 Assault on Precinct 13     30    20.04    Action
## 3 Pooh's Heffalump Movie     20    18.08 Adventure
## 4              Constantine     75    75.98    Action
## 5                  Hostage     75    34.64    Action
## 6                   Robots     80   128.20 Adventure
```

```
nrow(movies)
```

```
## [1] 35
```

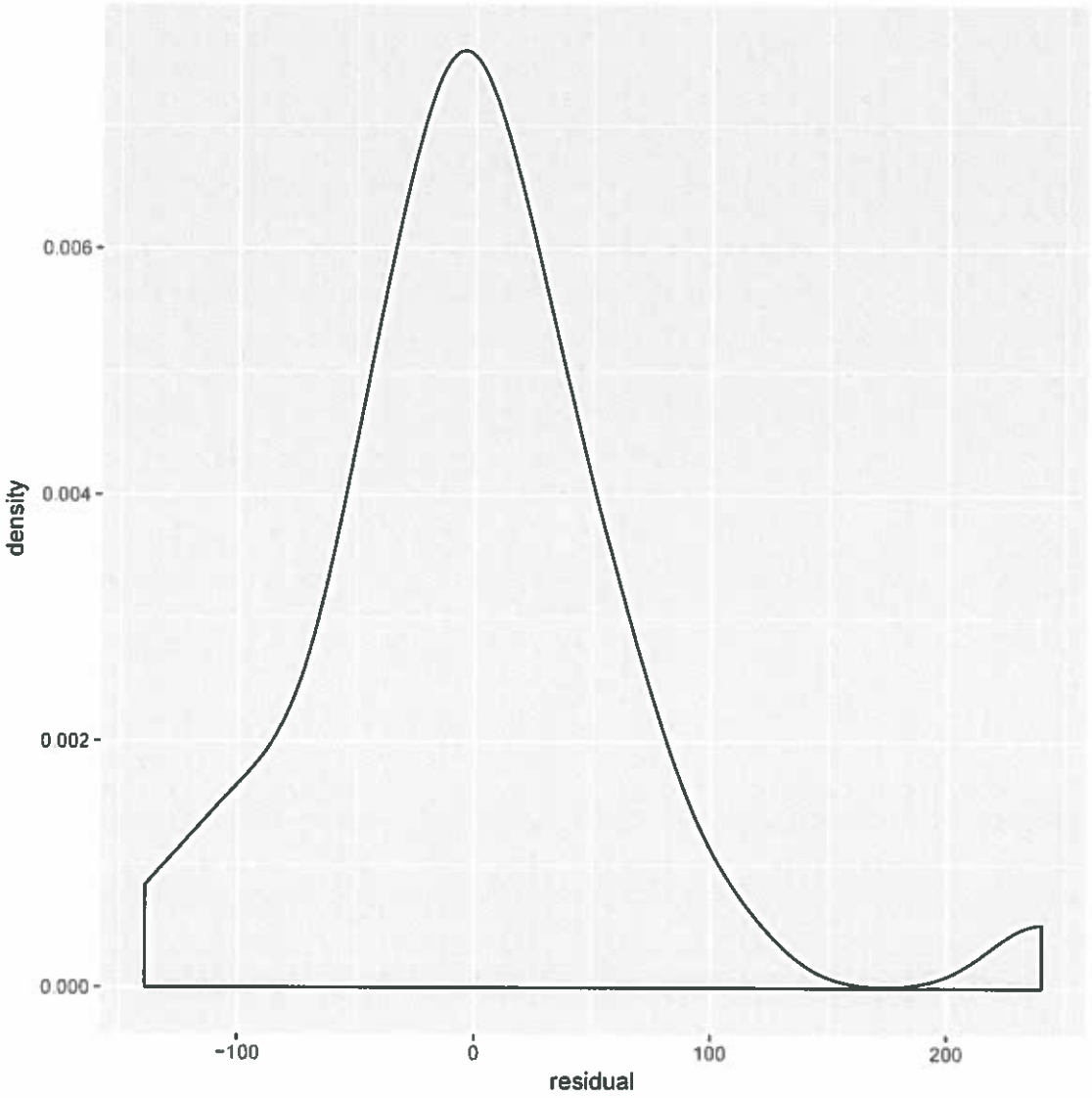
```
ggplot() +
  geom_point(mapping = aes(x = budget, y = earnings), data = movies)
```



```
## A model fit
earnings_model <- lm(earnings ~ budget, data = movies)
summary(earnings_model)

##
## Call:
## lm(formula = earnings ~ budget, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -139.02  -36.17   -5.18   30.78  240.21
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -15.314     22.273   -0.69      0.5
## budget         1.351      0.278    4.85  2.8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.2 on 33 degrees of freedom
## Multiple R-squared:  0.416,Adjusted R-squared:  0.399
## F-statistic: 23.5 on 1 and 33 DF,  p-value: 2.85e-05


## A plot of residuals
movies <- mutate(movies, residual = residuals(earnings_model))
ggplot() +
  geom_density(mapping = aes(x = residual), data = movies)
```



(a) Check the assumptions for the linear model. Extra Credit: If any assumptions are violated, suggest something you could do to address those limitations.

Outliers: there are outliers.
Linear relationship: yes, satisfied
Independent: I'm not sure about this. For example, maybe some of these movies have the same lead actor and they might have similar performance?
To proceed, we just have to assume that independence holds.
Normally distributed errors: This mostly holds, after the outlier we have already mentioned.
Equal spread of residuals: This does not hold, there is more variability in earnings as the movie's budget increases.
These problems could likely be fixed with a data transformation.
We really shouldn't proceed without doing a transformation first.

Regardless of your answer to part (a), let's proceed using the results from this model.

(b) (2 points) Write down the population model equation.

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

(c) (1 point) Write down the model equation again, filling in the estimated coefficients from the R output.

$\hat{Y}_i = -15.314 + 1.351 * X_i$

(d) (2 points) Interpret the model's slope coefficient in context.

For each increase of $1 million in a movie's budget, its predicted earnings increases by $1.351 million.

(e) (2 points) The budget for "Wallace Gromit: The Curse of the Were-Rabbit" (classified as an Adventure movie) was 30 million dollars. Based on this model, what is the predicted earnings for this movie?

$\hat{y}_i = -15.314 + 1.351 * 30 = 25.216$

$25.216 million.

(f) (2 points) The actual earnings for "Wallace Gromit: The Curse of the Were-Rabbit" was 56 million dollars. What is the residual for this movie?

$e_i = y_i - \hat{y}_i = 56 - 25.216 = 30.784$

(g) (2 points) Give the value of the residual standard deviation and its interpretation, using the 95 part of the 68-95-99.7 rule.

The residual standard deviation is 68.2.

For about 95% of the movies in this sample, the predicted earnings was within $\pm 2 * 68.2$, or $\pm$ 136.4, of the actual earnings.

(h) (2 points) Give the value of R^2 for this model and its interpretation.

$R^2 = 0.416$

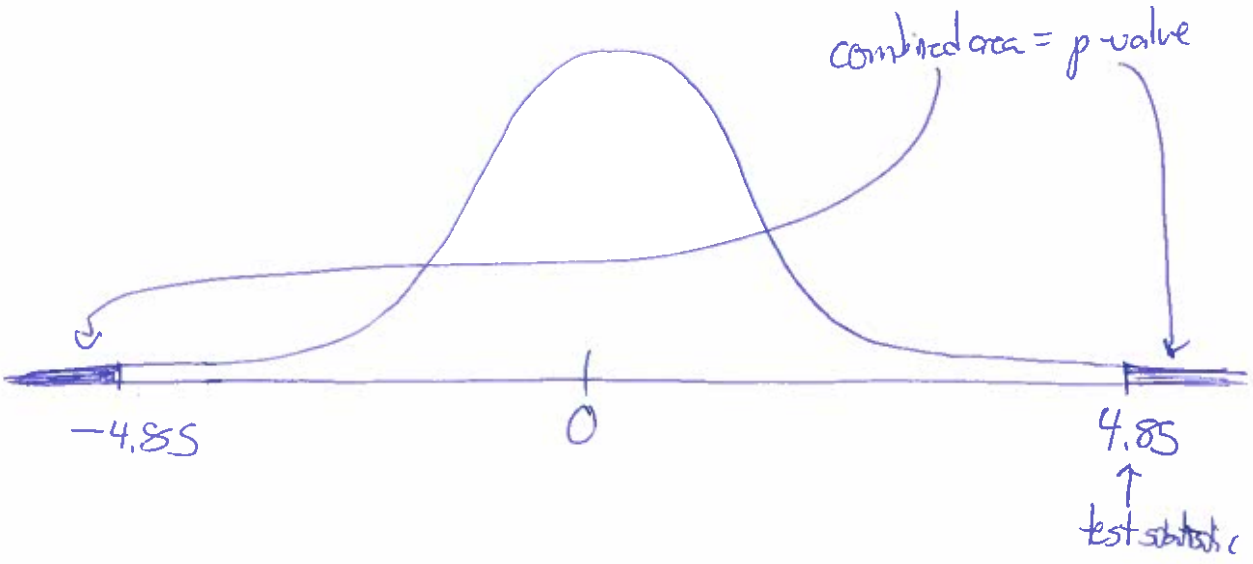
This linear model accounted for about 41.6% of the variation in movie earnings.

(i) (6 points) Is there a statistically significant linear association between movie budget and earnings? State a null and alternative hypothesis in terms of the coefficients you set up in part (b), write down the p-value from the R output, and draw a conclusion in the context of the problem.

$H_0: \beta_1 = 0$ vs. $H_A: \beta_1 \neq 0$.

The p-value is $2.8 * 10^{-5}$. The p-value is less than commonly used significance cut-offs such as $\alpha = 0.05$, so we can reject the null hypothesis. The data offer enough evidence to conclude that the slope describing the relationship between budget and earnings is not equal to 0 in the population of all movies.

(j) (3 points) Draw a picture of the sampling distribution that's relevant for the test you conducted in part (i), locate the value corresponding to the test statistic, and shade in the area corresponding to the p-value for the test.



(k) (6 points) Write down a 95% confidence interval for the coefficient in the model describing the relationship between a movie's budget and its earnings. Interpret the interval in context, including a discussion of what you mean by "95% confident". You will need to use one of the numbers in the following R output to construct your interval:

```
qt(0.95, df = 34)

## [1] 1.691

qt(0.95, df = 33)

## [1] 1.692

qt(0.975, df = 34)

## [1] 2.032

qt(0.975, df = 33)

## [1] 2.035
```

use 0.975 for a 95% C.I. and use df = n-2. Here the sample size is 35, so n-2 = 35-2 = 33.

$b_1 \pm t^* \cdot SE(b_1)$

$1.351 \pm 2.035 * 0.278$

$[0.785, 1.917]$

We are 95% confident that the slope of a line describing the relationship between budget and earnings in the population of all movies is between 0.785 and 1.917.

If we were to take many different samples and compute a different 95% C.I. based on each sample, about 95% of those confidence intervals would contain the true slope in the population.