Midterm Information

- In class on Tuesday (Oct 11): 1:30-2:45pm
- You are allowed a calculator (without internet access) and two pages of double-sided notes (4 sides total)
- The exam covers problem sets 0-4, lectures 1-10, and labs 1-5
- Practice exams are on Canvas
- Extra office hours are posted on Canvas (Monday and Tuesday)

Problem 1: Pride goeth before a fall (Just for fun)

Alex Roygant (A. Roygant) is in a class with 99 other people, and the class's only assignments are 11 weekly quizzes. At the end of the semester, the quiz grades for each student are averaged to get a final score. Alex brags that he scored in the 90^{th} percentile on each quiz. What is the lowest his final percentile could be in the class? Note: There are no assumptions made about scores being independent or scores being unique.

Alex's final percentile could be zero. There are 99 other students, and Alex scores in the 90^{th} percentile, so he scores better than 90 students on every exam, but there are up to 9 students who score better than him. Consider a case where 90 students get a 10, he gets an 11, and 9 students get 100. Now, on the next quiz, have 9 of the students who previously got 10s get 100s, have the people who got 100s get 10s, and have Alex get an 11 again. Repeat this 9 more times with a new group of 9 people each time. Then, Alex's average is an 11, but the average for every other student is $\frac{10 \cdot 10 + 100}{11} \approx 18.2$.

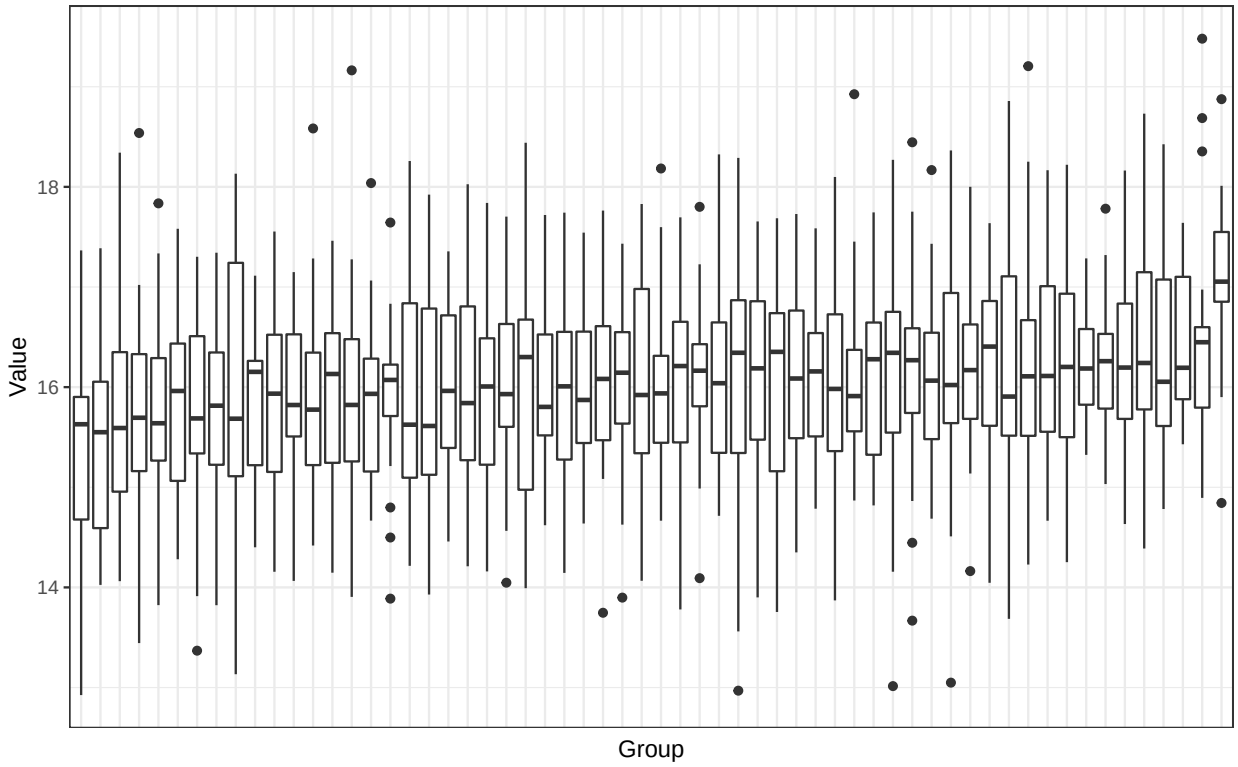
Problem 2: Statistical testing (Parts are unrelated unless otherwise specified)

(a) In June 2022, the New York Times reported that in the last year (June 2021 to June 2022) the per capita death rate from COVID-19 had become higher for white Americans than for Black and Hispanic Americans. However, many public health officials and statisticians pointed out that in every age bracket, Black and Hispanic Americans were still more likely than white Americans to die of COVID-19. How is this possible?

This is an example of Simpson's paradox. In the U.S., white people are considerably older (43.7 years) than Black (34.6 years) or Hispanic (29.8 years) people on average (Brookings 2019). Since COVID-19 deaths are much more common among older people, even though the death rates were higher for Black and Hispanic Americans in every age bracket, combining all the age groups indicated that more white people per capita were dying of Covid.

(b) Consider the following plot comparing 60 groups. Look carefully for groups that might be different from the rest.

```
library(ggplot2)
set.seed(33)
n = 20
sigma = 1
off_group <- rnorm(n, 17, sigma)
rest_groups <- rnorm(n * 59, 16, sigma)
data <- data.frame("Group" = c(rep(paste0("", 1:60), each = n)),
                   "Value" = c(rest_groups[1:(4*n)], off_group,
                               rest_groups[(4*n + 1):(59*n)]))
ggplot(data, aes(x=reorder(Group, Value), y=Value)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x=element_blank(), #remove x axis labels
        axis.ticks.x=element_blank(), #remove x axis ticks
        ) +
  xlab("Group")
```




```
summary(aov(Value~Group, data))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Group         59   68.9   1.168   1.163  0.191
## Residuals   1140 1145.1   1.004
```

```
pairwise_tests = as.numeric(unlist(pairwise.t.test(data$Value, as.factor(data$Group),
                                                   p.adjust.method = "none")))
pairwise_tests <- pairwise_tests[!is.na(pairwise_tests)]
mean(pairwise_tests < 0.05)
```

```
## [1] 0.07175141
```

i. Which (if any) of the ANOVA assumptions are violated?

The ANOVA assumptions are independence, equal variances within groups, and normality of the observations within groups. Equal variance and normality are satisfied. Independence cannot be checked without knowing the data generating process.

ii. Do you expect the F-test to be significant? Why or why not?

Even though there is one group that appears to have higher values than the rest, there are enough groups that this will probably be insignificant. (Note that the code to generate this data intentionally placed this group higher than the rest but that the test indeed came out insignificant.)

iii. If you ran pairwise t-tests for all of the groups, what proportion of the tests would you expect to be significant?

A) 0
B) 0.033
C) 0.050
D) 0.082

Assuming that the group on the right is almost always significantly different from the rest of the groups and that the rest of the groups have an $\alpha = 0.05$ probability of a false positive difference, the proportion of rejections will be about $\frac{59}{\binom{60}{2}} + 0.05\left(1 - \frac{59}{\binom{60}{2}}\right) = 0.082$

(c) A company is testing a new drug intended to reduce plaques of a misfolded protein in a rare disease. The company's biostatistics team has determined that for a preliminary study where all participants will be given the drug, they need a sample size n to achieve a Type 1 error rate of 0.05 and a Type 2 error rate of 0.2 for a t-test of H_0 : the drug has 0 effect on reducing plaques vs H_a : the drug has some effect on reducing plaques. However, when it comes time to enroll participants in the trial, the doctors can only find one large family where n people are affected by the disease. If the trial consists of just the n people from this family, provide plausible estimates of the Type 1 and Type 2 error rates and a brief explanation of why they will or will not change.

Type 1: 0.1, Type 2: 0.1. (Anything with Type 1 error above 0.05 and Type 2 error below 0.2 works.) The family members are likely to be more similar to one another than to a random person with the disease, so their ourcomes will likely be correlated. As we saw in Pset 1, when there is correlation in a sample, a t-test based on that sample is likely to overstate the true effect because we will underestimate the variance of the t-statistic. Therefore, we will be more likely to reject the null whether it is true or not, increasing our Type 1 error rate but decreasing our Type 2 error rate.

```
library(MASS)
nsims <- 10000
uncorrolated <- vector(length = nsims)
for (i in 1:nsims) {
  uncorrolated[i] <- t.test(rnorm(15, 0.1, 1))$p.value
}
corrolated <- vector(length = nsims)
Sigma <- matrix(0.5, 15, 15)
diag(Sigma) <- 1
for (i in 1:nsims) {
  corrolated[i] <- t.test(mvrnorm(1, rep(0.1, 15), Sigma))$p.value
}

mean(uncorrolated < 0.05)
```

```
## [1] 0.0622
```

```
mean(corrolated < 0.05)
```

```
## [1] 0.6
```

(d) Lotsa Cash is comparing incomes between Massachusetts and New Hampshire and decides to use a $\log_2$ transformation on the data because of its right skew. Lotsa obtains a confidence interval of (0.23, 0.66) for the difference in means (Massachusetts - New Hampshire) on the log scale. What should she conclude on the original scale? What assumption is required for this conclusion?

Assuming that the distributions of income are approximately symmetric on the log scale, the means will be about the same as the medians, and the medians are invariant to log transformation. Therefore, she should conclude that we are 95% confident the interval $(2^{0.23}, 2^{0.66}) = (1.17, 1.58)$ captures the true ratio of median incomes for Massachusetts vs New Hampshire.

(e) If 10 i.i.d. observations are generated from a Normal distribution with mean 0, which of the following have the t_9 distribution?

A) $\frac{\bar{X}}{\sqrt{\frac{S_X^2}{10}}}$

B) $\frac{\bar{X}}{\text{SE}(\bar{X})}$

C) $\frac{\bar{X}}{\frac{\text{UB}-\bar{X}}{t_9^*}}$ where UB is the upper bound in a 95% confidence interval for μ

D) $F_{t_9}^{-1}(U)$ where $F_{t_9}^{-1}$ is the t_9 quantile function and U is the p-value of a one-sample 2-sided t-test when the null is true.

All of these follow the t_9 distribution. The first is our usual t-statistic. The second is the usual t-statistic in standard error notation. The third uses the fact that $\frac{\text{UB}-\bar{X}}{t_9^*} = \text{SE}(\bar{X})$. The fourth uses the fact that the p-values are uniform under the null since the t_9 CDF is continuous and the fact that the quantile function is continuous for a t_9 distribution.

(f) For each variable, choose the transformation most likely to make it normal from the following list: log, exponential, square root, logit, reciprocal.

i. Proportion of people below the poverty line in US counties: Logit, used for proportions
ii. Population in US cities: Log, likely to very right skewed
iii. Area of individual napkins produced by a factory: Square root, each side length probably has a normal distribution so the area is probably a normal squared

(g) For each of the following scenarios, choose the 2-sample comparison from the following list most likely to fix any issues with the original data: unpooled t-test, paired t-test, log-transformed t-test, rank-sum test, permutation test.

i. The values in each group are bounded above by 5 and are left skewed with many points less than 0. You do not care about any particular statistic but rather about comparing the whole distributions. Rank-sum test: this is the only test that will compare the whole distributions, and it is robust to skew and location transformations.
ii. The values in each group are bounded above by 5 and are left skewed with many points less than 0. You want to compare the means of the groups. Permutation test for means: This test is robust to skew and shifts.
iii. The values in each group are measurements of the same thing at different times (one time in each group). You want to compare the means of the groups. Paired t-test: You have two times for each measurement, so there is a natural pairing.
iv. Each group has 80 observations, and the distribution of each group's values is slightly right skewed. You want to compare the means of the groups. Unpooled t-test: There are many observations, and if the data is not very skewed, the distributions of the sample mean will be approximately normal.

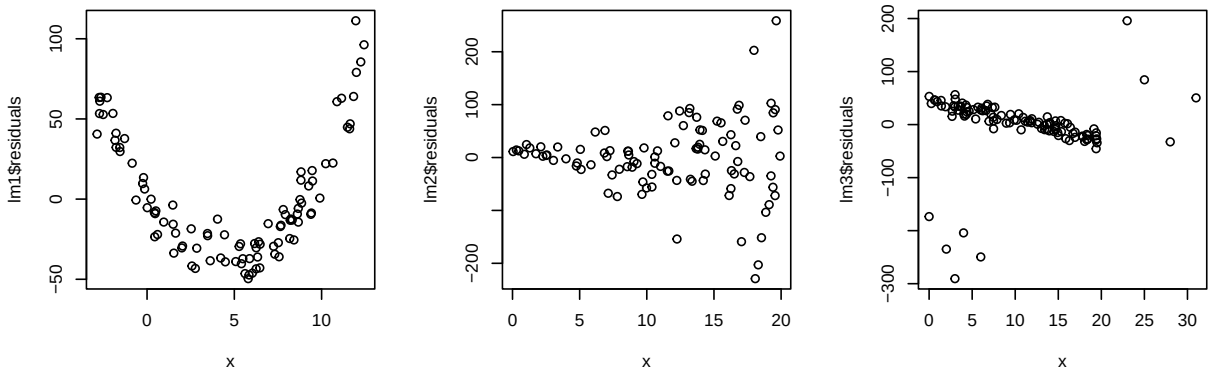
Problem 3: Regression (Parts are unrelated unless otherwise specified)

(a) In *Thinking Fast and Slow*, Daniel Kahneman recalls teaching Israeli fighter pilots the evidence-based idea that rewarding good performance is more effective than punishing poor performance. One of the experienced instructors disagreed, noting that when he praised a pilot for good performance the pilot rarely performed better on the next flight and when he criticized a pilot for poor performance the pilot often performed better on the next flight. Both Kahneman's teaching and the instructor's experience were likely true. How can this be?

This is a classic case of regression to the mean. When a pilot performs very poorly on one flight, the pilot will often perform better on the next just by chance. When a pilot performs unusually well on a flight, the pilot is unlikely to perform even better on the next. Kahneman was noting that in the long run (not as subject to regression to the mean), rewarding good performance tends to produce better results.

(b) Estimate the correlation of the residuals and the predictors in the following plots from simple linear regressions:

```
par(mfrow = c(1,3))
# Quadratic
n <- 100
x <- runif(n, -3, 13)
y <- (x-3)^2 * 2 + rnorm(n, 0, 10)
lm1 <- lm(y~x)
plot(lm1$residuals ~ x)

# Nonconstant variance
x <- runif(n, 0, 20)
y <- x * 2 + rnorm(n, 0, x*5)
lm2 <- lm(y~x)
plot(lm2$residuals ~ x)

x <- runif(n, 0, 20)
y <- 20 + x * 2 + rnorm(n, 0, 10)
x <- c(x, 23, 25, 28, 31, 0, 3, 2, 6, 4)
y <- c(y, 300, 200, 100, 200, -200, -300, -250, -242, -208)
lm3 <- lm(y~x)
plot(lm3$residuals ~ x)
```



0 for all. Residuals are always uncorrelated with the predictors in an OLS model with an intercept. If they were correlated, they would've been incorporated into the $\hat{\beta}$ prediction.

(c) Circle which of the following is not equivalent to the rest:

A) $\sum_{i=1}^{n}(X_i - \bar{X})\bar{X}$
B) $(\sum_{i=1}^{n} X_i^2) - n\bar{X}^2$
C) $\sum_{i=1}^{n}(X_i - \bar{X})X_i$
D) $\sum_{i=1}^{n}(X_i - \bar{X})^2$

A is not equivalent to the rest (and is in fact 0). Expanding the inside of the sum and pulling out terms that are not indexed quickly shows that B, C, and D are the same.

(d) The following summary is from a data set with the following columns:

- `epi_eh`: Environmental health policy score measuring how well a country is protecting its citizens from environmental health risks.
- `epi_cch`: Climate change score measuring progress to reduce pollutants including carbon dioxide, methane, and fluorinated gases.
- `mad_gdppc`: GDP per capita

Interpret the coefficients and significances of each predictor in the model.

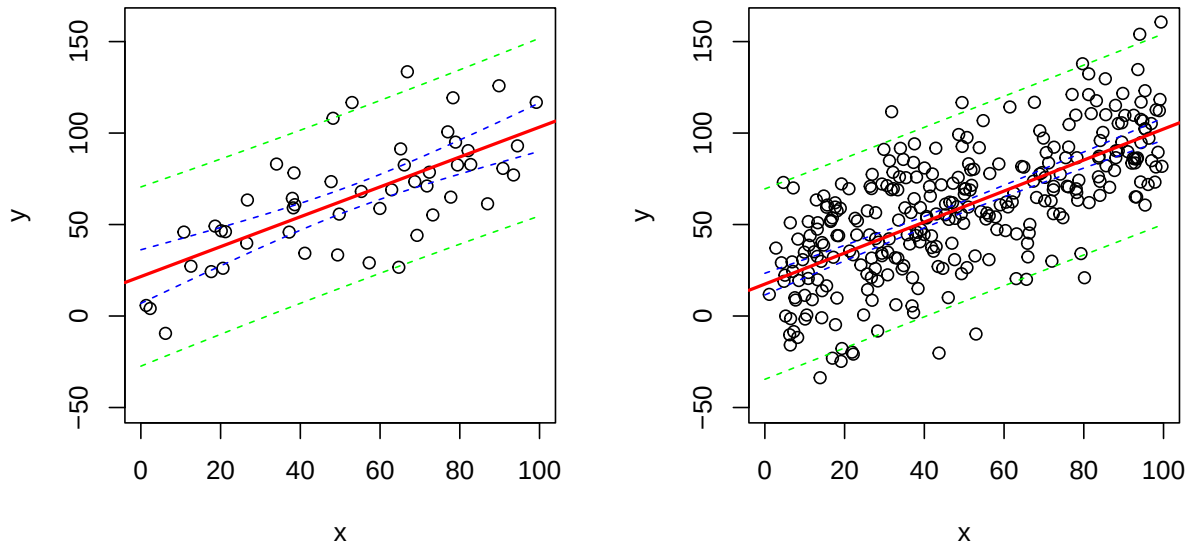
```
countries <- read.csv("data/countries.csv")
summary(lm(epi_eh ~ epi_cch + log(mad_gdppc, 2), countries))
```

```
##
## Call:
## lm(formula = epi_eh ~ epi_cch + log(mad_gdppc, 2), data = countries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.7246  -6.8701  -0.2615   8.3251  29.2407
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -94.15448    8.60541 -10.941  < 2e-16 ***
## epi_cch             0.43548    0.07876   5.529 1.32e-07 ***
## log(mad_gdppc, 2)   8.86217    0.79336  11.170  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.94 on 156 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
## F-statistic: 207.9 on 2 and 156 DF,  p-value: < 2.2e-16
```

Every 1 point increase in a country's climate change score is associated with a 0.44 point increase in its environmental health policy score after controlling for the association with GDP, and the result is very significant (t-stat 5.53, p-value $< 10^{-6}$). A doubling in a country's GDP is associated with a 8.86 point increase in its environmental health policy score after controlling for the association with its climate change score, and the result is very significant (t-stat 11.17, p-value $< 10^{-15}$).

(e) Two scatterplots with regression lines are shown below. Trace a 95% confidence and 95% prediction interval for each.

```r
set.seed(1)
par(mfrow = c(1, 2))

# Sparse plot
n <- 50
x <- runif(n, 0, 100)
y <- 20 + 0.8 * x + rnorm(n, 0, 25)
lm1 <- lm(y ~ x)

new.x = seq(0,100,0.5)
new.yhat = predict(lm1, newdata=data.frame(x = new.x),
                   interval = c("confidence"))
pred.yhat = predict(lm1, newdata=data.frame(x = new.x),
                    interval = c("prediction"))

plot(y~x, xlim=c(0, 100), ylim=c(-50, 160))
abline(lm1,col="red",lwd=2)
lines(new.x,new.yhat[,2], col="blue", lty=2)
lines(new.x,new.yhat[,3], col="blue", lty=2)
lines(new.x,pred.yhat[,2], col="green", lty=2)
lines(new.x,pred.yhat[,3], col="green", lty=2)

# Dense plot
n <- 300
x <- runif(n, 0, 100)
y <- 20 + 0.8 * x + rnorm(n, 0, 25)
lm2 <- lm(y ~ x)

new.yhat = predict(lm2, newdata=data.frame(x = new.x),
                   interval = c("confidence"))
pred.yhat = predict(lm2, newdata=data.frame(x = new.x),
                    interval = c("prediction"))
plot(y~x, xlim=c(0, 100), ylim=c(-50, 160))
abline(lm2,col="red",lwd=2)
lines(new.x,new.yhat[,2], col="blue", lty=2)
lines(new.x,new.yhat[,3], col="blue", lty=2)
lines(new.x,pred.yhat[,2], col="green", lty=2)
lines(new.x,pred.yhat[,3], col="green", lty=2)
```



The confidence interval is in blue and the prediction interval is in green. The prediction interval should be roughly the same between the two plots, but the confidence interval should be smaller in the second.

(f) Consider a data set with a continuous variable Y and a categorical variable X with equal proportions of 0s and 1s (n each). Exactly one of the following tests gives a different p-value. Which one? Assume the sample variances of each group are not the same.

A) Unpooled t-test
B) t-test for $\beta_1 = 0$ in the linear model `lm(y ~ x)`
C) Overall regression f-test for the resulting model of `lm(y ~ x)` having any predictive ability
D) Contrast test with $C^T = [0 \ 1]$ and $\vec{\beta}^T = [\beta_0 \ \beta_1]$ for the linear model `lm(y ~ x)`

```
contrast.test <- function(fit_lm, vec1, vec2) {
  beta.hat = coef(fit_lm)
  C = vec1 - vec2
  t.stat = C %*% beta.hat/sqrt(t(C) %*% vcov(fit_lm) %*% C)
  p.value = 2*(1-pt(abs(t.stat),df=fit_lm$df.residual))
  return (c("t.stat" = t.stat, "p.value" = p.value, "df" = fit_lm$df.residual))
}

x <- rep(c(0,1), each=10)
y <- 10 * x + rnorm(20, 0, 5)
fit_lm <- lm(y~x)
lm_sum <- summary(fit_lm)

c("t-test" = t.test(y~as.factor(x), alternative="two.sided")$p.value,
  "lm" = lm_sum$coefficients[2,4],
  "f-test" = pf(lm_sum$fstatistic[1],lm_sum$fstatistic[2],
                lm_sum$fstatistic[3],lower.tail=FALSE),
  "contrast" = contrast.test(fit_lm, c(0, 1), c(0, 0))[2])
```

```
##           t-test               lm     f-test.value contrast.p.value
##     7.381499e-05     4.337804e-05     4.337804e-05     4.337804e-05
```

The degrees of freedom in the unpooled t-test will be $n-1$ or with Satterthwaite:

$$\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}=\frac{\left(\frac{s_1^2+s_2^2}{n}\right)^2}{\left(\frac{s_1^4+s_2^4}{n^2(n-1)}\right)^2}=(n-1)\frac{s_1^4+2s_1^2s_2^2+s_2^4}{s_1^4+s_2^4}=(n-1)\left(1+\frac{2s_1^2s_2^2}{s_1^4+s_2^4}\right)\leq 2(n-1)$$

with equality only when the variances are the same. However, because the problem description required the variances of each group to be different, the degrees of freedom do not equal $2n-2$. Since the other tests use a t_{n-2} null distribution, the unpooled t-test will give a different result. Some algebra can show that the test statistic T equals the test statistic for the slope from the linear regression as long as there are equal numbers of observations from both groups. (This is also shown in the example code, or you can just use the fact that the other three statistics are equal and use the same distribution to conclude that the degrees of freedom difference doesn't balance out some test statistic difference.) Note that if we had used a pooled t-test, all the p-values would have been the same.

The f-test uses the statistic:

$$F=\frac{\frac{\text{SSM}}{\text{df}_M}}{\frac{\text{SSE}}{\text{df}_E}}=\frac{\frac{\sum_{i=1}^n(\hat{Y}_i-\bar{Y})^2}{1}}{\hat{\sigma}^2}=\frac{\sum_{i=1}^n(\hat{\beta}_1(X_i-\bar{X}))^2}{\hat{\sigma}^2}=\left(\frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/\text{SSX}}}\right)^2$$

with 1 and $n-2$ degrees of freedom. If $T\sim t_{n-2}$, $T^2\sim F_{1,n-2}$, so $P(-|t_{\text{obs}}|\leq T\leq|t_{\text{obs}}|)=P(T^2\leq t_{\text{obs}}^2)=P(F\leq t_{\text{obs}}^2)$. Also, as seen above, $f_{\text{obs}}=t_{\text{obs}}^2$. Therefore, the p-value is the same as in a t test for slope.

The t-test from the linear model uses the sampling distribution of $\hat{\beta}_1$ for its t-statistic:

$$\frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{1/\text{SSX}}}$$

The contrast test uses the statistic

$$T=\frac{\vec{C}^T\hat{\vec{\beta}}}{\hat{\sigma}\sqrt{C^T(X^TX)^{-1}C}}=\frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{1/\text{SSX}}}$$

($C^T(X^TX)^{-1}C=\frac{1}{\text{SSX}}$ can be seen from Pset 4.)

An easier way to find which p-value is different is to note that all four tests are testing the same null hypothesis of $\hat{\beta}_1=0$, and three are based on the linear model and one is based on the t-test. The three tests based on the same linear model that are testing the same hypothesis should give the same p-values or our system is broken.

(g) In a last ditch effort to draw conclusions from her data, Auda Fidese has decided to test every combination of her five β coefficients for significance in contrast tests. Specifically, she is going to test all $\vec{C}^T$ of the form $[0\ \ I_1\ \ I_2\ \ I_3\ \ I_4\ \ I_5]$ where I_j is an indicator of β_j being in the test. Assume the overall model F-test null is true and she is rejecting tests where the p-value is below α.

i. Find the expected number of Type 1 errors in this setup.

There are 31 possible tests because each of the 5 β_j can be included or not included, but at least one needs to be included, so we throw out the all-0 case. The probability of making a Type 1 error on any individual test is α, so by linearity, the expected number of Type 1 errors is 31α.

ii. Explain why the probability of making any Type 1 error is not $1-(1-\alpha)^{31}$.

This would be correct if each of the contrast tests were independent, but they clearly are not. Not only can the $\hat{\beta}_j$ be correlated, many of the tests involve the exact same $\hat{\beta}_j$!

Problem 4: Ben O'Meal's bootstraps

Ben O'Meal is trying to create a 95% confidence interval for the proportion of nights he stays up past midnight. He has recorded $n = 30$ night's worth of data and found that he stayed up past midnight 18 times. In his sleep deprivation, Ben cannot remember the proper way to build a confidence interval from this data, so he decides to use a studentized bootstrap.

(a) Show the correct way to build such a confidence interval, commenting on any assumptions necessary.

Ben should use a z-based confidence interval. We are assuming his sleep times are independent each night, which is probably reasonable. He stayed up past midnight 18 times and went to sleep before midnight 12 times, so the regularity condition (>10) is satisfied, and we can use a Normal approximation to the binomial. The confidence interval would be

$$(\hat{p} - z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = (0.6 - 1.96 \cdot 0.089, 0.6 + 1.96 \cdot 0.089) = (0.425, 0.775)$$

(b) Find the confidence interval Ben will produce, and show that it's wider than the proper confidence interval.

When he resamples, Ben will be drawing from a $\text{Bin}(n, \hat{p})$, so his resampled variance will be $\frac{\hat{p}(1-\hat{p})}{n}$. Then, his confidence interval will be

$$(\hat{p} - t^*_{29}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}, \hat{p} + t^*_{29}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}) = (0.6 - 2.0452 \cdot 0.091, 0.6 + 2.0452 \cdot 0.091) = (0.414, 0.786)$$

which is indeed slightly wider than before.

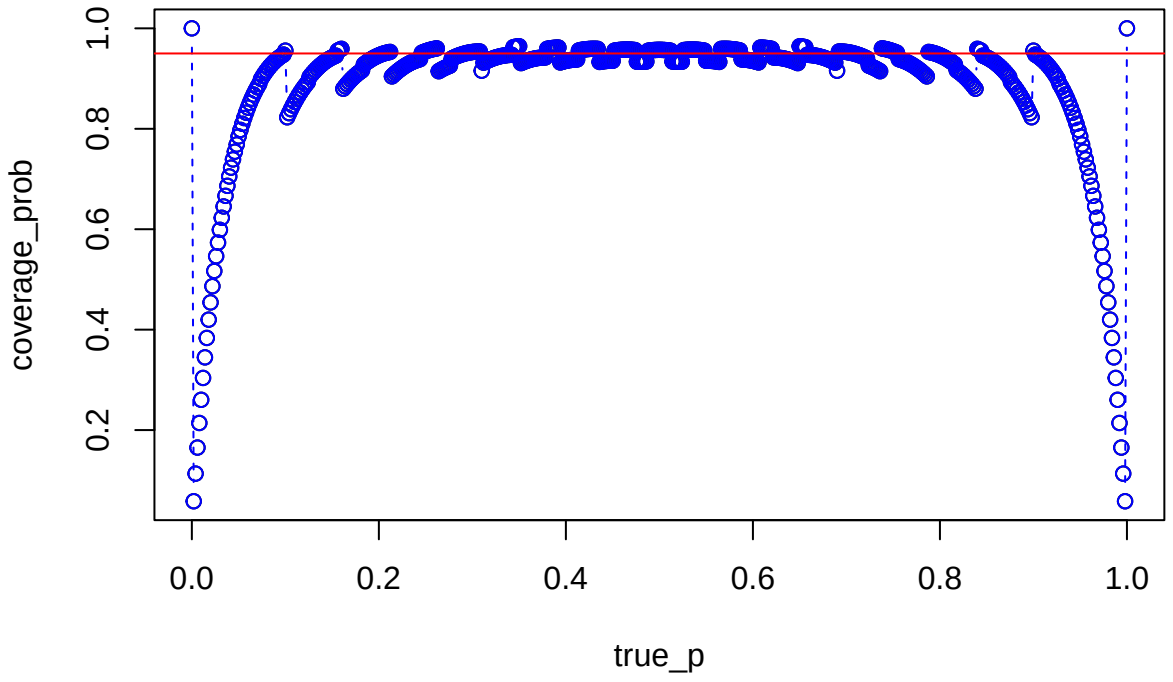
(c) Find the exact probability that a confidence interval created from Ben's method will capture the true p. Your answer may be left in terms of indicator functions and one sum.

If we know $\hat{p}$, we know whether the interval will capture p, so we just need the probability of observing a $\hat{p}$ that will create a capturing interval.

$$\begin{aligned} P(\text{Capture}) &= P\left(|p - \hat{p}| \le t^*_{29}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right) \\ &= \sum_{k=0}^{n} I\left(|p - \hat{p}| \le t^*_{29}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \Big| \hat{p} = \frac{k}{n}\right) P\left(\hat{p} = \frac{k}{n}\right) \\ &= \sum_{k=0}^{n} I\left(|p - \hat{p}| \le t^*_{29}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \Big| \hat{p} = \frac{k}{n}\right) \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

The coverage probability can be seen in the plot below for different p.

```
n <- 30
k <- 0:n
phat <- k/n
true_p <- seq(0, 1, 0.002)
coverage_prob = vector(length=length(true_p))
for (i in 1:length(true_p)) {
  p = true_p[i]
  coverage_prob[i] = sum(ifelse(abs(phat - p) <= qt(0.975, n-1) * sqrt(phat*(1-phat)/(n-1)),
                          choose(n,k) * p^k * (1-p)^(n-k), 0))
}
plot(true_p, coverage_prob)
lines(true_p, coverage_prob, col = "blue", type = "b", lty = 2)
abline(h=0.95, col="red")
```


coverage_prob
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
true_p

Problem 5: Recruiting the right Baller

A small liberal arts college outside of Boston believes that their secret to basketball success is a statistician under the pseudonym Coach t. Coach t knows of two identical twins in the area who are high school juniors, Alice Baller (A. Baller) and Nota Baller. One of the twins is very good at basketball, but the other is not very good. Coach t just received 20 games of film from the Baller family, but the Baller family neglected to label whether the film was of Alice or Nota. Coach t is too embarrassed to ask which twin is in the film, but she knows that Alice scores Pois(18) points a game while Nota scores Pois(9). Coach t only has time to watch one game, and she figures the family would only send film of Alice, so she decides that unless the twin makes 12 or fewer points in the first game she'll add all the games to Alice's history. (Otherwise she'll add it to Nota's.)

(a) Find the probability of making a Type 1 and a Type 2 error with this plan. You may leave your answer as a single sum.

A Type 1 error occurs if Alice scores 12 or fewer points. Using the PMF of a Poisson, this occurs with probability

$$\sum_{k=0}^{12} \frac{e^{-18}18^k}{k!} \approx 0.092$$

A Type 2 error occurs if Nota scores 13 or more points. This happens with probability

$$1 - \sum_{k=0}^{12} \frac{e^{-9}9^k}{k!} \approx 0.124$$

(b) Coach t gets a text from her friend Student t saying he's going to miss their scheduled lunch because he's too hungover on account of his "job" that he never speaks about. Coach t decides to use this extra time to watch another game and now decides that she'll add the games to Alice's history if the twin makes more than 12 points averaged across the two games. Assume that the points made in each game are independent. What are the probabilities of Type 1 and Type 2 errors now?

Making more than 12 points averaged across the games is equivalent to making 25 or more points total. The sum of independent Poissons is Poisson of the sum of their parameters, so the number of points Alice makes in both games combined has the distribution Pois(36). The new probability of a Type 1 error is then

$$\sum_{k=0}^{24} \frac{e^{-36}36^k}{k!} \approx 0.022$$

Likewise, the probability of a Type 2 error is

$$1 - \sum_{k=0}^{24} \frac{e^{-18}18^k}{k!} \approx 0.068$$

(c) Using a Normal approximation to the Poisson, show that every additional game Coach t watches will reduce both the chance of a Type 1 and the chance of a Type 2 error. (Note that a Normal approximation to a Poisson does not hold generally, but here we are essentially using a scaled Poisson.)

The total points made in n games is Pois($18n$) for Alice and Pois($9n$) for Nota. Because the mean and variance of the Poisson are the same, these are approximately $\mathcal{N}(18n, 18n)$ and $\mathcal{N}(9n, 9n)$. A z-statistic for Alice scoring $12n$ or fewer points is

$$\frac{12n - 18n}{\sqrt{18n}} = -\sqrt{2n}$$

so the z-statistic decreases with increasing n, and it becomes increasingly unlikely to observe a value less than the z-statistic under the null. Likewise, for Nota, the z-statistic for scoring more than $12n$ points is

$$\frac{12n - 9n}{\sqrt{9n}} = \sqrt{n}$$

This increases with n, so it becomes increasingly unlikely for Nota to score more than $12n$ points.

Problem 6: Imka Fused and her mixed up data

Imka Fused is scrambling to finish her final project in Stat 931. She has a data frame with two columns, and she has run a simple linear regression to predict one from the other. Unfortunately, she mixed up the column names and can't remember which is the predictor and which is the response. The deadline is minutes away and she doesn't have time to figure out which is which, but she still wants whatever she writes to be correct. Help Imka make some true claims about her data regardless of whether she used the model `response ~ predictor` or the model `predictor ~ response`.

(a) Show that the R^2s of both models are the same.

In a simple linear regression, $R^2 = r^2$ where r is the coefficient of correlation for the predictor. R^2 is the same for both models because r is symmetric for Y and X:

$$r = \frac{1}{n-1}\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$$

(b) Show that the overall model F-statistics are the same and that the F-test gives the same p-value in both models.

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSM}}{\text{SST}}$$

Therefore, $\frac{\text{SSE}}{\text{SST}}$ and $\frac{\text{SSM}}{\text{SST}}$ are equal between the models. Then, $\frac{\text{SSM}}{\text{SSE}} = \frac{\text{SSM}}{\text{SST}} / \frac{\text{SSE}}{\text{SST}}$ is equal in both models. The F-statistics are given by

$$F = \frac{\text{SSM}/\text{df}_M}{\text{SSE}/\text{df}_E} = \frac{\text{SSM}}{\text{SSE}/\text{df}_E}$$

and df_E is equal in both models, so the F statistics are equal. Also, the null distribution for both F tests is F_{1,df_E}, so the p-values of the F-tests are the same.

(c) Show that the t-statistic for a test of $\beta_1 = 0$ is the same in both models and that the p-values of the tests are the same.

The t-statistic is:

$$\begin{aligned} T &= \frac{\frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}}{\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}} \\ &= \frac{\frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}}{\sqrt{(1-R^2)\text{SST}/\text{df}_E}} \\ &= \frac{\frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}}{\sqrt{(1-R^2)\sum_{i=1}^n (Y_i - \bar{Y})^2/\text{df}_E}} \\ &= \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\frac{1-R^2}{\text{df}_E}(\sum_{i=1}^n (X_i - \bar{X})^2)(\sum_{i=1}^n (Y_i - \bar{Y})^2)}} \end{aligned}$$

This is symmetric in X and Y, so the t-statistics are the same for both models. Additionally, the t-statistic has the T_{n-2} distribution under the null in both models, so the p-values are the same.

(d) Briefly show that the $\hat{\beta}_1$ and $\text{SE}(\hat{\beta}_1)$ of each model need not be the same.

$$\hat{\beta}_{1,y\sim x} = \frac{r_{XY} S_Y}{S_X}$$

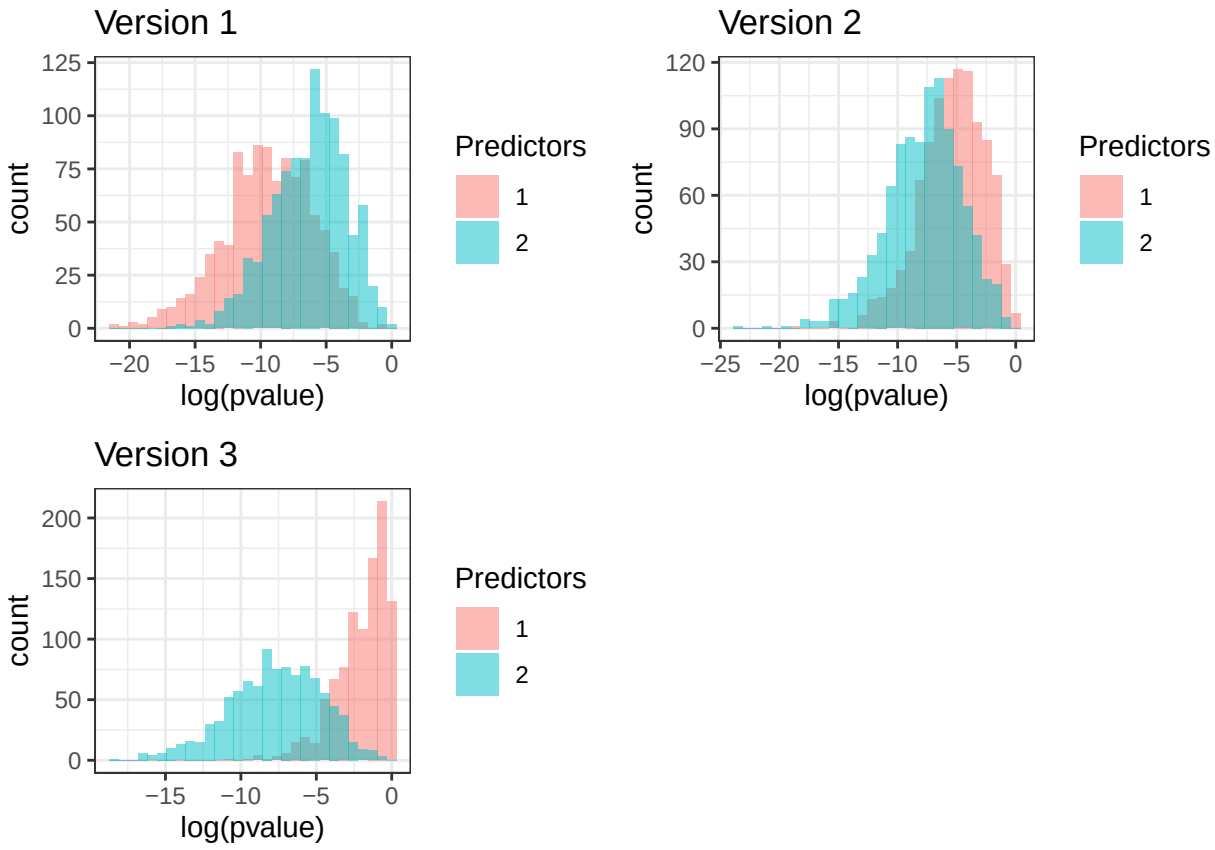
which is not symmetric in X and Y, so it will only be the same in the two models if $S_X = S_Y$. Additionally, the t-statistics are equal, so $\frac{\hat{\beta}_{1,y\sim x}}{\text{SE}(\hat{\beta}_{1,y\sim x})} = \frac{\hat{\beta}_{1,x\sim y}}{\text{SE}(\hat{\beta}_{1,x\sim y})}$, so the standard errors need not be the same unless the slope estimates are the same.

Problem 7: To add or not to add (predictors)

Read through the code for the following simulation and summarize what it is showing.

```
library(ggpubr)
set.seed(139)

nsims = 1000
n = 20
beta_1 = 2
beta_2 = 2
sigma = 2

run_simulation <- function(Sigma) {
  # Model with predictors for only X_1
  pvals_single = vector(length = nsims)
  for (i in 1:nsims) {
    x = mvrnorm(n = n, rep(0, 2), Sigma)
    y = x %*% c(beta_1, beta_2) + rnorm(n, 0, sigma)
    pvals_single[i] = summary(lm(y ~ x[,1]))$coefficients[2, 4]
  }

  # Model with predictors for X_1 and X_2
  pvals_double = vector(length = nsims)
  for (i in 1:nsims) {
    x = mvrnorm(n = n, rep(0, 2), Sigma)
    y = x %*% c(beta_1, beta_2) + rnorm(n, 0, sigma)
    pvals_double[i] = summary(lm(y ~ x))$coefficients[2, 4]
  }

  out_data = data.frame(Predictors = as.factor(c(rep(1, nsims), rep(2, nsims))),
                        pvalue = c(pvals_single, pvals_double))

  return(out_data)
}

# Version 1
Sigma = cbind(c(1, 0.5), c(0.5, 1))

out_data <- run_simulation(Sigma)

plot1 <- ggplot(out_data, aes(x = log(pvalue), fill = Predictors)) +
  geom_histogram(alpha=0.5, position="identity", bins=30) +
  theme_bw() +
  ggtitle("Version 1")

# Version 2
Sigma = cbind(c(1, 0), c(0, 1))

out_data <- run_simulation(Sigma)

plot2 <- ggplot(out_data, aes(x = log(pvalue), fill = Predictors)) +
  geom_histogram(alpha=0.5, position="identity", bins=30) +
  theme_bw() +
```

```
  ggtitle("Version 2")

# Version 3
Sigma = cbind(c(1, 0), c(0, 10))

out_data <- run_simulation(Sigma)

plot3 <- ggplot(out_data, aes(x = log(pvalue), fill = Predictors)) +
  geom_histogram(alpha=0.5, position="identity", bins=30) +
  theme_bw() +
  ggtitle("Version 3")

ggarrange(plot1, plot2, plot3, ncol = 2, nrow = 2)
```



The simulation is demonstrating how including extra predictors in a multiple regression model can increase or decrease the significance of a predictor relative to a model with only that one predictor. When a new predictor is added that is significantly collinear with the predictor already present, the original predictor loses significance (Version 1). In the context of a test for $\beta_1 = 0$, we can imagine a contrast vector $\vec{C}^T = [0\ 1\ 0]$ and the test statistic

$$\frac{\vec{C}^T\hat{\vec{\beta}}}{\hat{\sigma}\sqrt{C^T(X^TX)^{-1}C}}$$

Increasing collinearity increases the diagonals of $(X^TX)^{-1}$, causing $\hat{\sigma}\sqrt{C^T(X^TX)^{-1}C}$ to be larger, decreasing the t-statistic and making the p-value higher. When a new predictor is added that is mostly noncollinear with the predictor already present and explains some of the residual variance, $\hat{\sigma}$ decreases in

$\hat{\sigma}\sqrt{C^T(X^TX)^{-1}C}$, making the t-statistic larger and the p-value smaller (Version 2). When this new predictor explains even more of the variance, $\hat{\sigma}$ is further reduced, and the t-statistic becomes even larger (Version 3).