hadCET <- read.csv("/var/folders/lx/7p7y6jl52hn43jc6f16jphzr0000gn/T//RtmprP4NvQ/data180f56f149c71", sep="")

dec.year <- hadCET$dec.year

temp <- hadCET$temp

## Define the sinusoid and cosine covariates

sin.term <- sin(2 \* pi \* dec.year)

cos.term <- cos(2 \* pi \* dec.year)

dec.year.sq <- dec.year^2

dec.year.cu <- dec.year^3

dec.year.qu <- dec.year^4

dec.year.pe <- dec.year^5

## Fit a regression model including the sin and cos terms, and a

model1 <- lm(temp ~ sin.term + cos.term + dec.year+ dec.year.sq+ dec.year.cu+ dec.year.qu+ dec.year.pe)

## Calculate the residuals and fitted values.

res <- resid(model1)

fits <- fitted(model1)

plot(fits, res, xlab="Fitted values", ylab="Residuals", main='Residuals on Fitted Values')

abline(h=0, lty=2)

plot(dec.year, res, xlab="Year", ylab="Residuals", main='Residuals on Decimal Year')

abline(h=0, lty=2)

hist(res, xlab="Residuals", main='Histogram of Residuals')

qqnorm(res)

qqline(res)



A. While the residuals do appear centered around the zero in the first two plots, their variance across all the fitted values is not consitent. There is greater variance in residuals for smaller fitted values. As the fitted values approach their median, the variance in the residuals appears to decrease, and post median, the variance appears to be increasing slightly. Lastly, similar to the example in class, when the temperature is low (low fitted value), the residuals can be abnormally low (approaching -6). The opposite extreme doesn't occur, for the residuals at low temps only approach 4.5. This phenomenon is also seen in the Histogram's leftward skewdness.

# Two

## Summarize the model

summary(model1)

B. As the sin term increases by one unit, the estimated change in the average temperature is positive 1.116e+04. As the cos term increases by one unit, the estimated change in the average temperature is negative 8.076e-01. It is important to note that these two terms fluctuate between -1 and 1. So if, for example, the sin term increased by one half, the associated average temperature increase is one half of 1.116e+04. Since the p-values given in the table are nearly zero, we can confirm that these terms contribute significantly to the prediction power of the model. The estimated coefficients for the dec.year, dec.year.sq, dec.year.cu,and dec.year.qu are more straighforward. They represent the estimated change in the average temperature as they increase by one unit. So as the square of the decimal year increased by one, the assoaciated average temperature increases by 1.969e-02. Since the p-values for these four terms in the table are so large, we can confirm that these terms do not contribute significantly to the prediction power of the model.The dec.year.pe coefficient's, however, were not calculated This is because dec.year.pe contained redunant information aka information that was already provided through some linear combination of the other terms mentioned above.

C. The Akaike information criterion (AIC) is a means of measuring the quality of models predicting the same response. It has no interpretable value and is useful only by means of comparison of one model to another. The AIC is equivalent to -2(log-likelihood) +2k where k is the number of parameters in the model. The log-likelihood reflects the overall fit of the model. Smaller values indicate worse fit. It takes residuals, number of terms and the amount of data points into account.

#ii

AIC(model1)

The AIC of the current model with enough terms to capture a 5th degree polynomail in the year is 11436.06. The AIC for the model including just the sin, cos, and dec.year terms is larger at 11467.72, and the AIC for the 3rd degree polynomial is smaller at 11434.87. Based on the AIC criterion, the 3rd degree polynomial model is the optimal model because it has the lowest AIC value.

(D).

## Fit a regression model including the sin and cos terms, and a

## linear function of 'dec.year'.

model2 <- lm(temp ~ sin.term + cos.term + dec.year + dec.year.sq + dec.year.cu)

#Run an F-Test where null model is third degree polynomial and alt model is fifth degree polynomial

RSS.null = sum(resid(model2)^2)

df.null = model2$df.residual

RSS.alt = sum(resid(model1)^2)

df.alt = model1$df.residual

Fstat = ((RSS.null - RSS.alt)/(df.null - df.alt)) / (RSS.alt/df.alt)

pf(Fstat, df1=1, df2=3185, lower.tail = F)

The F statistic for the F test where the null model is a third degree polynomial and alt model is fifth degree polynomial is 0.8053612. The corresponding p-val is 0.3695631. This means that the alternate model, or in this case the model of fifth degree, does not add significantly more value than the model under the null which, here, is the model of third degree polynomial.

Question 2.

(A).

#calc sample p

sp=318/1191

#95% C.I z stat

zstat=qnorm(1-0.05/2)

lower.b = sp - (zstat)\*sqrt((sp\*(1-sp))/1191)

upper.b = sp + (zstat)\*sqrt((sp\*(1-sp))/1191)

When repeating this process/experiment, we can expect the true population proportion to be within the confidence interval calculated via the data 95% of the time. The confidence interval for this iteration of the experiment is (0.2418778, 0.2921272). If this is one of the 95 out of 100 times, the sample proportion is in fact within these two numbers. Else it is part of the 5 out of 100, and the population proportion is outside the numbers.

(B).

#calc sample p with two successes and two failures

spW=320/1195

#95% C.I z stat

zstat=qnorm(1-0.05/2)

sepW = sqrt((spW\*(1-spW))/1195)

lower.bW = sp - (zstat)\*sepW

upper.bW = sp + (zstat)\*sepW

The confidence interval calculated using Wilson's method of adding two successes and two failures is (0.2418967, 0.2921084). Given that the numbers involved are relatively large, the interval did not change much. This interval smaller and thus less conservative than the interval above calculated via plug-in method. This makes sense because adding additional trials adds more data and allows us to be more confidence in statistical inferences.

(C).

#participation rate

pr=1195/1236

The participation rate for this study was 96.68285%. The fact that they did not respond might be dependent on the typical length of their sermons. If, for example, a congregations do not answer the question because they know they have too long of a sermons, our sample p will tend to be smaller than the true population p.

(D).

People are predictably irrational. They are filled with biases and often those biases skew their perception (done both conciously and unconciously) of reality. I would bet that, in this scenario, ministers, priests, rabbis, and other staff persons or leaders would tend to underestimate how long their sermons are. For many people, attending a sermon isn't exactly the most stimulating activity. In order to feel better about themselves, a religious leader might say their sermons are shorter than they really are to look and feel more appealing. Also, for many religious leaders, time probably flys while they are giving their sermons. This too, would cause them to beleive their sermons are shorter than they really are. Given these too potential facts, a sample proportion of how many sermons are more than 30 minutes would be below what the true population proportion is. These potential biases need to be taken into account when we decide how confident we are in our results and potential discoveries.

3.

(A).

p, the parameter of interest for this problem can be defined as the population proportion of potatoes that have major defects in the truck of 3,000.

(B).

In order to prove that the normal approximation is valid for this test, we must show that the expected number of successes, np, is greater than or equal to 10 and that the expected number of failures, n(1-p), is greater than or equal to 10. In this example, np = (150)(0.1) = 15 and n(1-p) = 150(1-0.1) = 135. Since these two evaluate above 10, we are valid in using normal approximation. We also need to show that the population size is at least 10 times as large as the sample to prove the number of successes (count of potatoes with major defects) in the SRS does in fact follow a Bernoulli distribution. 150 \* 10 >= 3000, so this constraint holds.

(C).

#phat

p = 0.1

#standard error of phat

pSe = sqrt((p\*(1-p))/150)

#Test Statistic

z=((8/150)-0.1)/(pSe)

#Get p val

pnorm(z)

(D).

The test statistic is -1.905159 and the p-val is 0.02837972. This says that, given p is in fact 0.10, the likelihood of getting data similar to what we got is 2.837972%. It is far more likely that p is less than 0.1, so in this case we reject the null hypothesis in favor of the fact that the population proportion of defecive potatoes of the 3,000 in the truckload is less than 0.1. So the inspector should be okay with the shipment going out of its station.