Final exam, STATS 401 F18

Name: **UMID**:

- You have a time allowance of 120 minutes. The exam is closed book and closed notes. Any electronic devices (including calculators) in your possession must be turned off and remain in a bag on the floor.
- If you need extra paper, please number the pages and put your name and UMID on each page.
- Responses will be assessed on quality of explanation as well as whether they lead to a correct answer.
- You may use the following formulas. Proper use of these formulas may involve making appropriate definitions of the necessary quantities.

1. $\mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}, \qquad \hat{\boldsymbol{\beta}} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}, \qquad \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}$
2. $\mathrm{Var}(X) = \mathrm{E}\big[(X - \mathrm{E}[X])^2\big] = \mathrm{E}[X^2] - \big(\mathrm{E}[X]\big)^2$
3. $\mathrm{Cov}(X, Y) = \mathrm{E}\big[\big(X - \mathrm{E}[X]\big)\big(Y - \mathrm{E}[Y]\big)\big] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$
4. $\mathrm{Var}(\mathbb{A}\mathbf{Y}) = \mathbb{A}\mathrm{Var}(\mathbf{Y})\mathbb{A}^{\mathrm{T}}, \qquad \mathrm{var}(\mathbb{X}\mathbb{A}^{\mathrm{T}}) = \mathbb{A}\,\mathrm{var}(\mathbb{X})\,\mathbb{A}^{\mathrm{T}}$
5. $(\mathbb{A}\mathbb{B})^{\mathrm{T}} = \mathbb{B}^{\mathrm{T}}\mathbb{A}^{\mathrm{T}}, \quad (\mathbb{A}\mathbb{B})^{-1} = \mathbb{B}^{-1}\mathbb{A}^{-1}, \quad (\mathbb{A}^{\mathrm{T}})^{-1} = (\mathbb{A}^{-1})^{\mathrm{T}}, \quad (\mathbb{A}^{\mathrm{T}})^{\mathrm{T}} = \mathbb{A}.$
6. The probability density function of the standard normal distribution is $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$
7. If a random variable is normally distributed, the probability it falls within one standard deviation of the mean is 68%, within two standard deviations of the mean is 95%, and within three standard deviations of the mean is 99.7%.
8. $f = \dfrac{(\mathrm{RSS}_0 - \mathrm{RSS}_a)/(q-p)}{\mathrm{RSS}_a/(n-q)}$.

Problem	Points	Your Score
1	5	
2	12	
3	3	
4	5	
5	7	
6	3	
7	3	
Total	38	

Acknowledgments: The climate data come from https://doi.org/10.1016/j.envsci.2012.03.008

The questions in this exam concern data you have seen previously in the course on global climate change from 1961 to 2010. Carbon dioxide (CO_2) levels in the atmosphere have been increasingly steadily, as recorded by the measurements taken at Mauna Loa observatory in Hawaii. An increasing trend in CO_2 matches increasing trends in both global economic activity and the global population, as well as many other socioeconomic phenomena. However, on shorter timescales, fluctuating geophysical processes such as volcanic activity and the El Nino Southern Oscillation (ENSO) may be important.

```
head(climate,3)
```

```
##   Year    CO2  GDP   Pop    ENSO Volcanic Emissions
## 1 1961 317.64 7.54 3.069 -0.2322   0.0024       9.5
## 2 1962 318.45 7.97 3.123 -0.7650   0.0024       9.8
## 3 1963 318.99 8.38 3.189 -0.1629   1.8454      10.4
```

- CO2: Mean annual concentration of atmospheric CO2 (parts per million by volume) at Mauna Loa.
- GDP: world gross domestic product reported by the World Bank.
- Pop: world population, in billions, reported by the World Bank.
- ENSO: an El Nino Southern Oscillation index from NOAA.
- Volcanic: an index of monthly estimated sulfate aerosols derived from NOAA.
- Emissions: estimated emissions of CO2 (million Kt) reported by the World Bank.

Throughout the exam, you may write y_i for the CO_2 concentration on the ith row of the data, t_i for the corresponding year, g_i for the world GDP, and m_i for the emissions index. The other variables will not be involved in models we consider here. We start by detrending the global CO_2 data, fitting a quadratic trend:

```
lm1 <- lm(CO2~Year+I(Year^2),data=climate)
summary(lm1)
```

```
##
## Call:
## lm(formula = CO2 ~ Year + I(Year^2), data = climate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10575 -0.54413 -0.00216  0.40201  1.56418
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.233e+04  1.993e+03   21.24   <2e-16 ***
## Year        -4.377e+01  2.008e+00  -21.80   <2e-16 ***
## I(Year^2)    1.140e-02  5.056e-04   22.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6655 on 47 degrees of freedom
## Multiple R-squared:  0.9991, Adjusted R-squared:  0.9991
## F-statistic: 2.625e+04 on 2 and 47 DF,  p-value: < 2.2e-16
```

1 *[5 points]*. Write down the probability model in subscript form for the quadratic trend fitted by `lm1` above. Include all details.

2 *[12 points].* Consider the F test from `summary(lm1)` above.

(a) *[2 points].* What are the null and alternative hypotheses for this F test? You can use notation that you defined in Question 1.

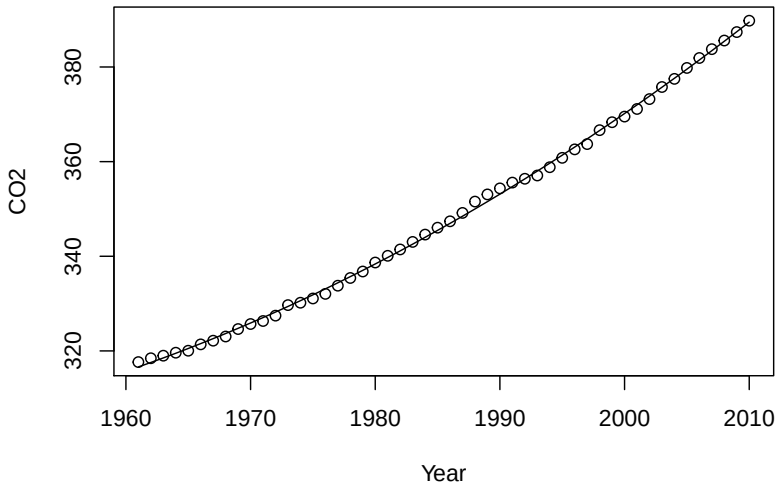
(b) *[4 points].* Explain the construction of the sample F-statistic in `summary(lm1)`. You should include explanation of why it is `on 2 and 47 DF`.

(c) *[4 points].* Explain how the p-value on the `F-statistic` line of the summary is calculated. Specifically, start by giving the general definition of a p-value and then explain how the definition applies to this specific test. Your explanation will likely include the phrases "sample test statistic", "model-generated test statistic", and "distribution under the null hypothesis".

(d) *[2 points].* What do you conclude from this F test?

The fitted values from `lm1` are plotted against time in the following plot:

```
plot(CO2~Year, data=climate)
lines(climate$Year,lm1$fitted)
```



3 *[3 points].* Looking at the fitted value plot, do you think the model is a good fit? Also, describe two additional analyses you would carry out to diagose misspecification in model `lm1`.

4 *[5 points].* Consider the `multiple R-squared statistic` from `summary(lm1)` above.

(a) *[1 point].* Describe how this R^2 statistic is calculated. Your answer can use notation, including sums of squares, that you have defined in answers to previous questions.

(b) *[2 points].* To what extent do you agree with the statement: "Because the multiple R-squared statistic is close to 1, the model fits well." Explain what you can and cannot generally conclude from a high R^2 about model specification, and then put this in the context of the specific analysis.

(c) *[2 points].* Explain why it may be preferable to look at the `Adjusted R-squared` statistic rather than the `Multiple R-squared` for some particular purpose.

5 *[7 points].* This question concerns a prediction interval. To work with prediction intervals, it is useful to first write down the model in matrix form.

(a) *[3 points].* Define $\mathbb{X}$ and $\boldsymbol{\beta}$ such that the probability model you wrote in Question 1 has matrix form $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. You can assume that $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ for a suitable choice of n.

(b) *[4 points].* Explain how to find a 95% prediction interval for CO_2 in 2018 using the data and the probability model fitted by `lm1`. You can use mathematical notation developed in part (a). A strong answer should demonstrate understanding of what a 95% prediction interval is and how it is constructed.

Hint: You will likely want to find a row vector $\mathbf{x}^*$ such that $Y^* = \mathbf{x}^*\boldsymbol{\beta} + \epsilon^*$ is a random variable modeling a new measurement with a new measurement error ϵ^* independent of $\epsilon_1, \ldots, \epsilon_n$. You may also use in your solution the notation $\hat{Y}^* = \mathbf{x}^*\hat{\boldsymbol{\beta}}$ for a model-generated fitted value at $\mathbf{x}^*$,

Now we can study the relationships between the detrended variables. Here, we are just going to look at global CO_2 , GDP and emissions. Rather than detrending with a quadratic model, we will detrend using the local linear regression function `loess()` which was found to work well in the health economics example in the notes. For the current purposes, we don't need to understand details about how `loess()` works. We put a 'd' in front of each detrended variable name in the following code.

```
climate$dCO2 <- resid(loess(CO2~Year,data=climate,span=0.5))
climate$dGDP <- resid(loess(GDP~Year,data=climate,span=0.5))
climate$dEmissions <- resid(loess(Emissions~Year,data=climate,span=0.5))
lm2 <- lm(dCO2~dGDP+dEmissions,data=climate)
summary(lm2)$coef
```

```
##               Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) 0.01254947 0.04945508 0.2537549 0.8007918
## dGDP        0.30173830 0.18793113 1.6055791 0.1150654
## dEmissions  0.21229065 0.15310970 1.3865265 0.1721283
```

```
lm3 <- lm(dCO2~dEmissions,data=climate)
summary(lm3)$coef
```

```
##               Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 0.01760106 0.05015955 0.3509015 0.727197709
## dEmissions  0.38387257 0.11143351 3.4448577 0.001195971
```

6 *[3 points]*. The coefficient for detrended global CO_2 emissions has much higher statistical significance in `lm3` than `lm2`. Explain this fact.

7 *[3 points]*. To what extent does the analysis above demonstrate that a major cause of fluctuations around the trend in global CO_2 levels is fluctuations in CO_2 emissions?