# Final exam, STATS 401 F18

**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 $i$th 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.

**Solution**:

$$Y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \epsilon_i, \quad i = 1, \ldots, n$$

where $n = 50$ and $\epsilon_1, \ldots, \epsilon_n \sim$ iid normal$(0, \sigma)$. Here, $Y_i$ models $CO_2$ levels for year $i$, and $\beta_0, \beta_1, \beta_2$ are unknown constants.

**Summary of grade scheme**: 1 point for the covariates $(1, t_i, t_i^2)$. 1 point for definition of $\epsilon_i$. 1 point for $\boldsymbol{\beta}$, with explanation. 1 point for $Y_i$. 1 point for $n$.

**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.

**Solution**:

$H_0$ : The probability model from Question 1 holds with $\beta_1 = 0$ and $\beta_2 = 0$.

$H_a$ : The probability model from Question 1 holds with $\beta_1$ and $\beta_2$ unconstrained.

**Summary of grade scheme**: 1 point for each hypothesis

(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`.

**Solution**:

Let $\text{RSS}_a$ be the residual sum of squares under $H_a$, that is, $\text{RSS}_a = \sum_{i=1}^n e_i^2$ where $e_1, \ldots e_n$ are the residuals from fitting

$$y_i = b_0 + b_1 t_i + b_2 t_i^2 + e_i, \quad i = 1, \ldots, n$$

with $b_0$, $b_1$ and $b_2$ chosen by least squares. $H_a$ fits 3 parameters so has $q = 50 - 3 = 47$ residual degrees of freedom.

Let $\text{RSS}_0$ be the residual sum of squares under $H_0$, that is, fitting

$$y_i = b_0 + e_i, \quad i = 1, \ldots, n$$

with $b_0$ chosen by least squares. $H_0$ fits 1 parameter so has $p = 50 - 1 = 49$ residual degrees of freedom.

The sample F statistic is

$$f = \frac{(\text{RSS}_0 - \text{RSS}_a)/(p - q)}{\text{RSS}_a/q}$$

with degrees of freedom being $p - q = 2$ and $q = 47$.

**Summary of grade scheme**: 1 point explaining each of $\text{RSS}_a$ and $\text{RSS}_0$, either in words or as a formula. 1 point for the formula for the sample f statistic. 1 point for explaining the 2 and 47 in terms of residual degrees of freedom

(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".

**Solution**:

The p-value is the probability that a model-generated test statistic takes a more extreme value than the sample test statistic under the null hypothesis. In this case, pval $= \text{P}(F > f)$ where $f$ is the sample F

statistic from 2(b) and $F$ is a model-generated F statistic under the probability model corresponding to the hypothesis $H_0$. $F$ is a random variable having the F-distribution on 2 and 47 degrees of freedom.

**Summary of grade scheme**: 2 points for the general definition (1 point if showing some comprehension but also a non-trivial error). 1 point for identifying the sample test statistic, 1 point for the model-generated distribution under $H_0$.

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

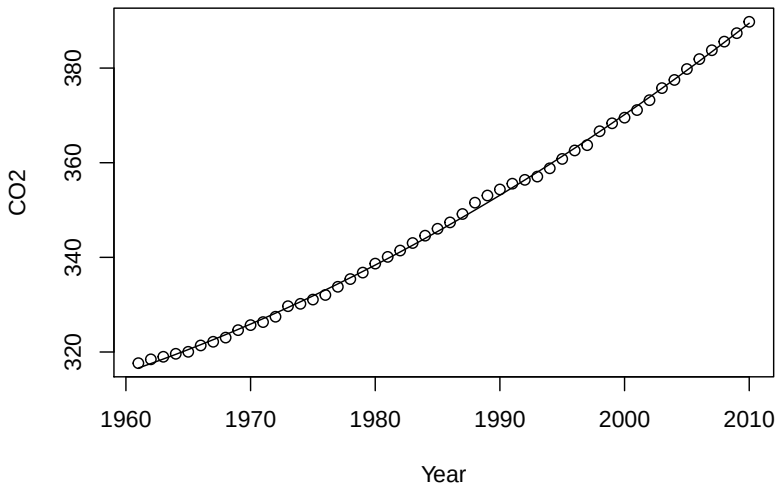
**Solution**:

The p-value is lower than any reasonable significance level. We clearly reject the null hypothesis and conclude that a constant expected value plus measurement error is not a good model for the data. In this case, the conclusion is not surprising. We are fitting the model in order to estimate the trend and subtract it. We are not particulary interested in the possibility that there is actually no trend.

**Summary of grade scheme**: 1 point rejecting. 1 point for some sensible interpretation of what this means in the context of this particular scientific analysis.

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`.

**Solution**:

The data follow the curve approximately, but we can nevertheless see long sequences of measurements all above or below the fitted line. This is inconsistent with the independent measurement errors in our probability model. Superficially, the curve seems like a good fit (and credit was given for noticing this) but "fit" formally means how well the data are modeled by the proposed probability model.

To see this pattern more clearly, we could make (i) a time plot of the residuals; (ii) a lag plot (plotting $e_i$ against the $e_{i-1}$); (iii) look for improved model fit for a different model (e.g., try including $t_i^3$ in the model and see if we obtain a statistically significant coefficient).

**Summary of grade scheme**: 1 point for a sensible comment about model fit; $2 \times 1$ points for two sensible additional analyses. The analyses should be clearly described, so just saying "residual plot" is not sufficient.

**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.

**Solution**:

$R^2 = 1 - \text{RSS}_a/\text{RSS}_0$ using the notation of 2(b).

**Summary of grade scheme**: 1 point for the definition.

(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.

**Solution**:

The $R^2$ statistic close to 1 says that the model fits much better than a simple constant mean model. It does not say whether or not there is room for additional model improvement. In situations like this where there is a clear trend, $R^2$ will always be high even when the model is significantly flawed, so $R^2$ is not a good way to assess if the model fits well.

**Summary of grade scheme**: 1 point for something sensible about what you can learn from $R^2$, 1 point for something sensible about a model fitting issue you can't learn from it.

(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.

**Solution**:

The unadjusted $R^2$ statistic cannot decrease when an additional explanatory variable is added to a linear model. This makes it hard to use comparison of $R^2$ values for model selection. Adjusted $R^2$ divides $\text{RSS}_a$ and $\text{RSS}_0$ by the corresponding residual degrees of freedom, penalizing $\text{RSS}_a$ for its additional parameters.

**Summary of grade scheme**: Maximum 2 of the following. 1 point for "model selection". 1 point for "$R^2$ doesn't decrease". 1 point for showing knowledge of the definitions of adjusted $R^2$ and $R^2$. 1 point for any other relevant and correct comment.

**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 $n$.

**Solution**:

Set $\mathbb{X}_{n\times 3} = [\ \mathbf{1}\ \ \mathbf{t}\ \ \mathbf{t}^2\ ]$ with $\mathbf{1} = (1, \ldots, 1)$, $\mathbf{t} = (t_1, \ldots, t_n)$ and $\mathbf{t}^2 = (t_1^2, \ldots, t_n^2)$. Then, set $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ in the notation of Question 1.

**Summary of grade scheme**: 1 point for each of $\mathbb{X}$ and $\boldsymbol{\beta}$. It is acceptable to write out $\mathbb{X}$ using .... 1 extra point if notation is correctly used (e.g., underscore for vectors, blackboard bold for matrices) and is consistent with notation used earlier.

(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 $\epsilon^*$ 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 $\mathbf{x}^*$,

**Solution**:

The sample version in matrix form is

$$\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e},$$

where $\mathbf{b} = (b_0, b_1, b_2)$ is a vector of least squares coefficients. Now, set $\mathbf{x}^* = (1, 2018, 2018^2)$. Using the notation in the hint, a 95% prediction interval is an interval

$$[\hat{y}^* - 1.96\text{SE}_{\text{pred}}, \hat{y}^* + 1.96\text{SE}_{\text{pred}}]$$

constructed so that $Y^*$ falls within

$$[\hat{Y}^* - 1.96\text{SD}_{\text{pred}}, \hat{Y}^* + 1.96\text{SD}_{\text{pred}}]$$

with probability 0.95, where $\text{SE}_{\text{pred}}$ is an estimate of $\text{SD}_{\text{pred}}$. In words, if the model is right the 95% prediction interval at $\mathbf{x}^*$ covers a new observation at $\mathbf{x}^*$ with probability 0.95. The variance of $Y^* - \hat{Y}^*$ is

$$\text{SD}^2_{\text{pred}} = \sigma^2 \mathbf{x}^* \left(\mathbb{X}^{\text{T}}\mathbb{X}\right)^{-1} \mathbf{x}^{*\text{T}} + \sigma^2$$

and so we take

$$\text{SE}_{\text{pred}} = s\sqrt{\mathbf{x}^* \left(\mathbb{X}^{\text{T}}\mathbb{X}\right)^{-1} \mathbf{x}^{*\text{T}} + 1}.$$

**Summary of grade scheme**: 1 point for $\mathbf{x}^*$. 1 point for $\text{SE}_{\text{pred}}$. 1 point for putting together the prediction interval. 1 point for a correct explanation of what a prediction interval is, either in words or as a probability statement. No credit was assigned for specification of $s$

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.

**Solution**. This arises because `dCO2` and `dEmissions` have high sample correlation, so they are close to collinear. When both are in the model, neither is significant because the data can't determine if one should be chosen over the other. When detrended emissions is used as the only explanatory variable in the model, there is strong evidence for an association with fluctuations in $CO_2$.

Quite a few students discussed the order of the variables in the summary table. In a table of least squares coefficients and their standard errors, order is unimportant. In an ANOVA table, order can be important.

**Summary of grade scheme**: 1 point for mentioning collinearity. 1 point for describing how collinearity affects standard errors. 1 point for a relevant comment relating this back to the data analysis.

**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?

**Solution**. This is an observational study, so an association does not imply a causal relationship unless one can rule out the possibility of confounding variables, that is, variables which influence both the explanatory variable and the response. We know that fluctuations in emissions are correlated with fluctuations in GDP.

Therefore, any other activity associated with economic activity (cutting down rainforest; emissions of some other gas that interacts via atmspheric chemistry to lead to $CO_2$. The alternative confounding explanations don't seem very plausible, but we are not environmental experts. It is plausible to claim based on common knowledge that there are no substantial confounders for the role of emissions: most pathways by which the economy affects $CO_2$ will be via emissions.

**Summary of grade scheme**: At most 3 points from the following. 1 point for "observational". 1 point for "confounding". 1 point for a relevant comment explaining what this means in practice. 1 point for a plausible example.