STAT 425 Assignment 4

Due Tuesday, March 22, 11:59pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

Most relevant class notes: 4.1.Collinearity, 4.2.GLS, 4.3.TestFit, 5.1.Polynomial

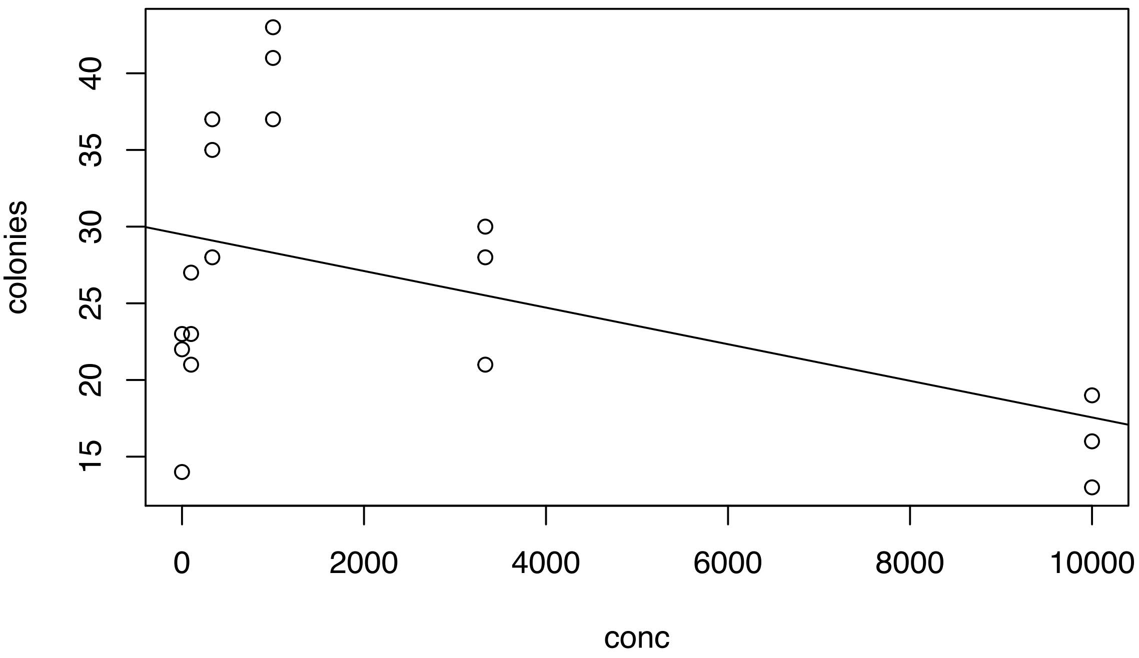
Problem 1

A study was conducted to see if a certain food dye (Acid Red 118) affected mutation rates in salmonella bacteria. The data are in the included file, "reddye.csv." The variables are concentration of red dye on the plate (`conc`), and number of mutation colonies developed on the plate (`colonies`).

a) Read the data into an R data frame, and make a scatterplot with `colonies` as the response and `conc` as the predictor. Include the least squares line on the graph. What does the slope of the LS line suggest about the relationship between `conc` and `colonies`? Does the line seem to fit the data?

Answer:

```r
data = read.csv("~/Desktop/STAT 425 TA spring/HW/hw4/reddye.csv")
attach(data)
plot_a = plot(conc, colonies) +
  abline(lm(colonies ~ conc))
```



The slope of the least squares line suggests that the relationship between conc and colonies is negative. The line seems to the fit the data fairly well, although the variance in colonies is higher when conc is lower.

b) Test the fit of the linear model in part a) versus the more general alternative model where conc is treated as a factor variable. Recall from class: if `conc` is treated as a factor variable, then we assume only that y has a different mean for each value of `conc` and nothing more. What do you conclude from the results?

Answer:

```
fit_a = lm(colonies ~ conc)
fit_b = lm(colonies ~ factor(conc))
anova(fit_a, fit_b)
```

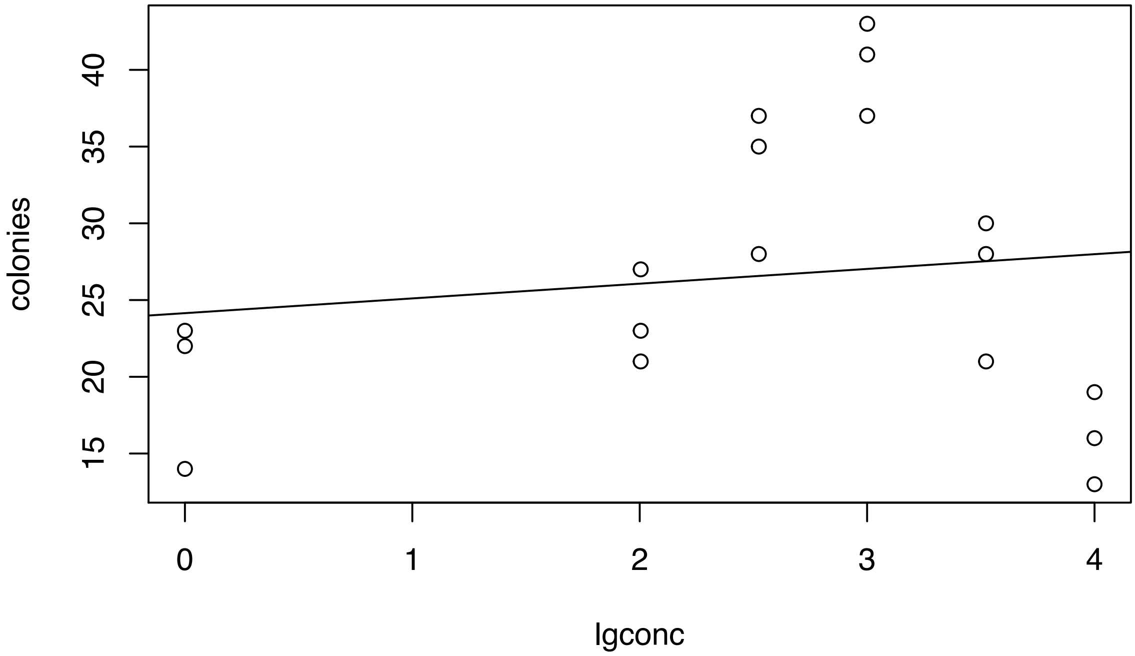
```
## Analysis of Variance Table
##
## Model 1: colonies ~ conc
## Model 2: colonies ~ factor(conc)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     16 1078.32
## 2     12  193.33  4    884.99 13.732 0.0001968 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 0.0001968, we can reject the null under any reasonable alpha and conclude that the model that treats conc as a factor variable is more suitable model.

c) Since the concentrations range over several orders of magnitude, a logarithmic transformation of `conc` might help. There is a problem, though, because the zero concentration would transform to $-\infty$. Instead, consider the constructed variable `lgconc=log10(1+conc)`. Make a scatterplot of `colonies` versus `lgconc`, including the least squares line for the corresponding linear regression model. What does the slope of the LS line suggest about the relationship between `lgconc` and `colonies`? Does the line seem to fit the data?

Answer:

```
lgconc = log10(1 + conc)
plot_c = plot(lgconc, colonies) +
  abline(lm(colonies ~ lgconc))
```



The slope of the least squares regression line suggests that the relationship between lgconc and colonies is fairly positive. The line does not seem to fit the data as the data takes on an almost parabolic shape rather than a linear one.

d) Test the fit of the model in c) by comparing it to the more general model that treats `lgconc` as a factor variable.

Answer:

```
fit_c = lm(colonies ~ lgconc)
fit_d = lm(colonies ~ factor(lgconc))
anova(fit_c, fit_d)
```

```
## Analysis of Variance Table
##
## Model 1: colonies ~ lgconc
## Model 2: colonies ~ factor(lgconc)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     16 1374.63
## 2     12  193.33  4    1181.3 18.331 4.765e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 4.765e-05, we have enough evidence to reject the null and assume that there is a lack of fit in the model in part c.

e) Obtain the R-square values and model F test p-values for three models:

$$
\begin{aligned}
\text{colonies} &= \beta_0 + \beta_1 \texttt{conc} + \texttt{error}\\
\text{colonies} &= \beta_0 + \beta_1 \texttt{lgconc} + \texttt{error}\\
\text{colonies} &= \beta_0 + \beta_1 \texttt{lgconc} + \beta_2 \texttt{lgconc}^2 + \texttt{error}
\end{aligned}
$$

Based on the results, which of these three models seems most reasonable and why? **Answer:**

```
fit_e1 = lm(colonies ~ conc)
fit_e2 = lm(colonies ~ lgconc)
fit_e3 = lm(colonies ~ lgconc + I(lgconc^2))

summary(fit_e1)
```

```
##
## Call:
## lm(formula = colonies ~ conc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.492  -5.920  -1.327   5.550  14.701
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.4919432  2.3530833  12.533 1.09e-09 ***
## conc        -0.0011932  0.0005441  -2.193   0.0434 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.209 on 16 degrees of freedom
## Multiple R-squared:  0.2311, Adjusted R-squared:  0.1831
## F-statistic: 4.809 on 1 and 16 DF,  p-value: 0.04343
```

```
summary(fit_e2)
```

```
##
## Call:
```

```
## lm(formula = colonies ~ lgconc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9885  -6.1654  -0.3378   6.9399  15.9719
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.1454     4.7664   5.066 0.000115 ***
## lgconc        0.9608     1.6887   0.569 0.577289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.269 on 16 degrees of freedom
## Multiple R-squared:  0.01983,    Adjusted R-squared:  -0.04143
## F-statistic: 0.3237 on 1 and 16 DF,  p-value: 0.5773
```

```
summary(fit_e3)
```

```
##
## Call:
## lm(formula = colonies ~ lgconc + I(lgconc^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7746  -4.7181   0.3789   4.6333  13.0402
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18.217      4.284   4.253 0.000695 ***
## lgconc        14.002      4.405   3.179 0.006230 **
## I(lgconc^2)   -3.362      1.080  -3.113 0.007133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.462 on 15 degrees of freedom
## Multiple R-squared:  0.4045, Adjusted R-squared:  0.3251
## F-statistic: 5.094 on 2 and 15 DF,  p-value: 0.0205
```

```
#Model 1: Multiple R squared value is .2311 and p-value is .04343
#Model 2: Multiple R squared value is .01983 and p-value is .5773
#Model 3: Multiple R squared value is .4045 and p-value is .0205
```

The third model seems to be the best fit out of the three because it has the lowest p-value at .0205 and the highest multiple R-squared value at .4045. A lower p-value suggests stronger evidence for a significant model and a higher multiple R-squared value means that the model's fitted values more closely resemble the response.

Problem 2

The `aatemp` data in the `faraway` library comes from the U.S. Historical Climatological Network. The data report annual mean temperatures in Ann Arbor Michigan for roughly 150 years.

a) Fit a linear trend model to temperature as a function of year and display the model summary. Does there appear to be a trend?

Answer:

```
library("faraway")
model = lm(temp ~ year, data = aatemp)
summary(model)
```

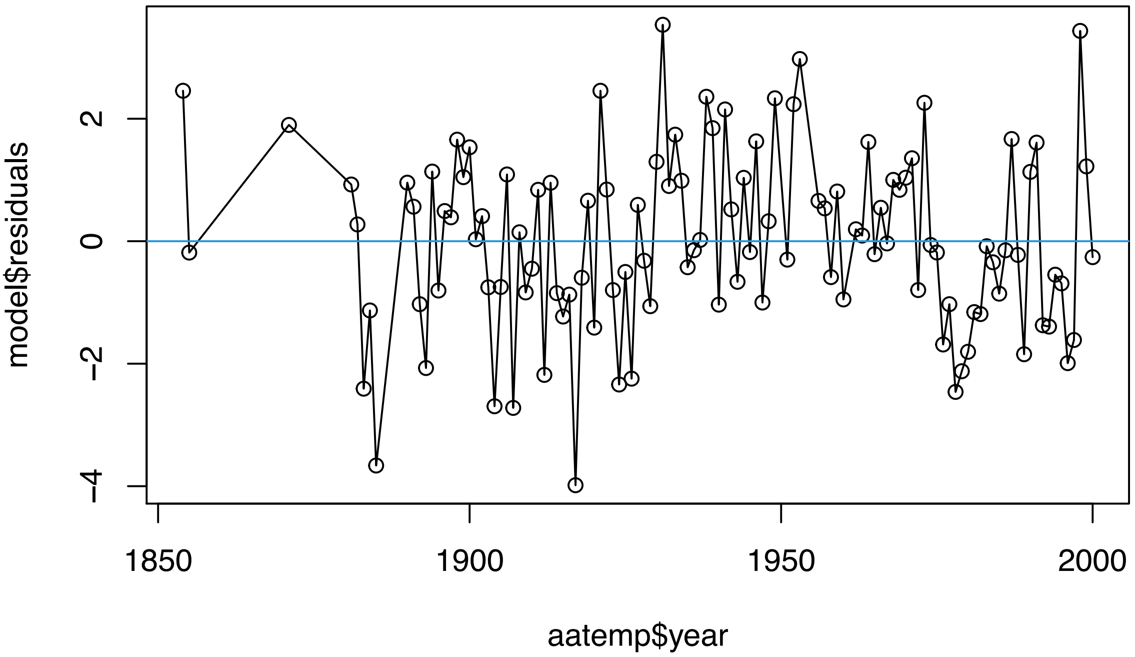
```
##
## Call:
## lm(formula = temp ~ year, data = aatemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9843 -0.9113 -0.0820  0.9946  3.5343
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.005510   7.310781   3.284  0.00136 **
## year         0.012237   0.003768   3.247  0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.466 on 113 degrees of freedom
## Multiple R-squared:  0.08536,    Adjusted R-squared:  0.07727
## F-statistic: 10.55 on 1 and 113 DF,  p-value: 0.001533
```

As we can see from the summary, p value for the slope coefficient is small, which implies there appears to be a trend.

b) For the model in a), plot residuals versus year, connecting the dots in the plot (`type='o'`), and adding a horizontal reference line at 0. Is there any evidence of serial correlation in the graph?

Answer:

```
plot(aatemp$year, model$residuals, type = 'o')
abline(h=0, col =4)
```



From the graph we can see the evidence of serial correction.

c) Based on the model in a), test for serial correlation between successive years using the Durbin-Watson test, and state your conclusion.

Answer:

```
library("lmtest")
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.6177, p-value = 0.01524
## alternative hypothesis: true autocorrelation is greater than 0
```

Since the p value is smaller than 0.05, we may reject the null hypothesis and conclude that there exists serial correction.

d) Using the `gls` function from the `nlme` library, fit a linear trend model using the AR1 form of correlation between years. Display the model summary. Does the trend line change much? How much correlation is there, based on the estimated AR1 correlation parameter?

Answer:

```
library("nlme")
armodel = gls(temp ~ year, correlation = corAR1(form = ~ year),
              data = aatemp)
summary(armodel)
```

```
## Generalized least squares fit by REML
##   Model: temp ~ year
##   Data: aatemp
##        AIC      BIC    logLik
##   426.5694 437.479 -209.2847
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~year
##  Parameter estimate(s):
##      Phi1
## 0.2303887
##
## Coefficients:
##                Value Std.Error  t-value p-value
## (Intercept) 25.18407  8.971864 2.807006  0.0059
## year         0.01164  0.004626 2.516015  0.0133
##
##  Correlation:
##      (Intr)
## year -1
##
## Standardized residuals:
##        Min         Q1         Med          Q3         Max
## -2.7230803 -0.6321970 -0.0520135  0.6645795  2.3775123
##
## Residual standard error: 1.475718
## Degrees of freedom: 115 total; 113 residual
```

The trend line does not seem changing much by checking the intercept and slope coefficients. Based on the model summary we can see the estimated AR1 correction is 0.2303887.

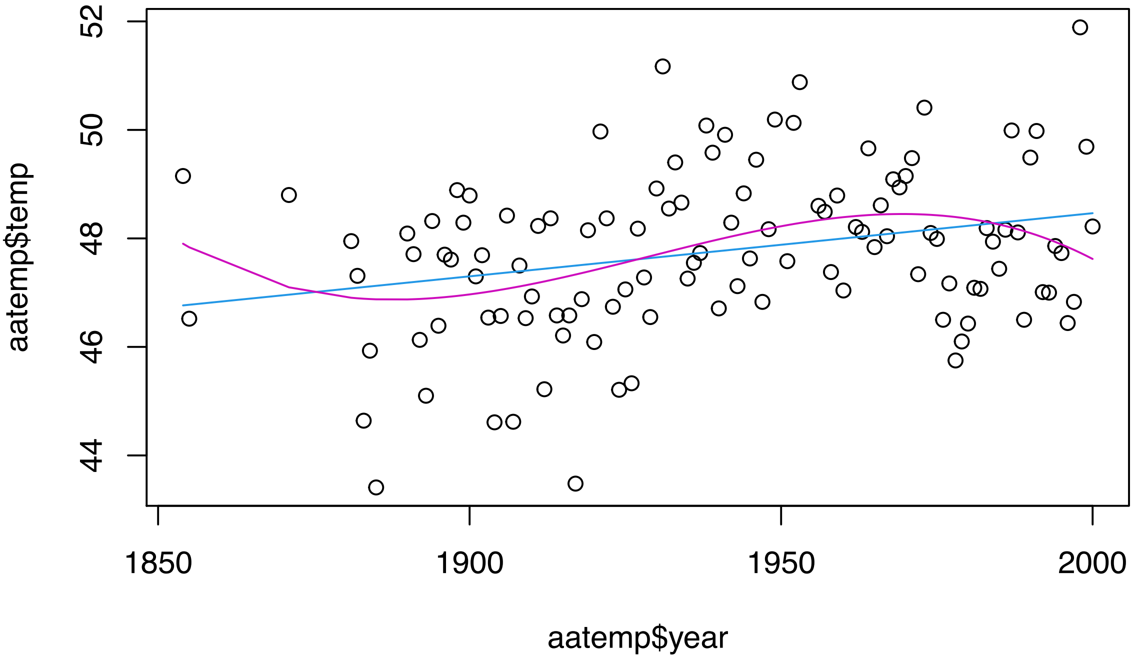
e) Again using the `gls` function, fit a cubic model (third order polynomial) for temperature as a function of year, with the AR1 form of correlation. Display the model summary. Make a scatter plot of `temp` versus `year`, and add the fitted curves from the linear and 3rd order polynomial models to the graph. One way to do this is with `lines` command after creating the plot, e.g. `lines(mod$fitted~year, data=aatemp)`. Which model seems to track the data better, based on what you see?

Answer:

```
armodel3 = gls(temp ~ year + I(year^2) + I(year^3),
               correlation = corAR1(form = ~ year), data = aatemp)
summary(armodel3)
```

```
## Generalized least squares fit by REML
##   Model: temp ~ year + I(year^2) + I(year^3)
##   Data: aatemp
##       AIC      BIC   logLik
##   466.35 482.6071 -227.175
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~year
##  Parameter estimate(s):
##      Phi1
## 0.214179
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 41531.81 20557.977  2.020228  0.0458
## year          -64.59    31.956 -2.021152  0.0457
## I(year^2)       0.03     0.017  2.023909  0.0454
## I(year^3)       0.00     0.000 -2.026180  0.0451
##
##  Correlation:
##           (Intr) year I(y^2)
## year      -1
## I(year^2)  1     -1
## I(year^3) -1      1   -1
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.6468228 -0.6658191 -0.1092529  0.7129614  2.8531996
##
## Residual standard error: 1.456044
## Degrees of freedom: 115 total; 111 residual
```

```
plot(aatemp$year, aatemp$temp)
lines(armodel$fitted ~ year, data = aatemp, col = 4)
lines(armodel3$fitted ~ year, data = aatemp, col = 6)
```



As we can see, the cubic one tracks the model much better.

Problem 3:

We delve into the theory for added variable plots and variance inflation factors. Consider a model of the form

$$\mathbf{y} = \mathbf{X}_0\mathbf{fi} + \mathbf{z}\gamma + \mathbf{e}$$

where $\mathbf{X_0}$ is $n \times p$ and full rank, $\mathbf{z}$ is $n \times 1$ and linearly independent of the columns of $\mathbf{X_0}$, $E(\mathbf{e}) = \mathbf{0}$, and $Cov(\mathbf{e}) = \sigma^2\mathbf{I}$.

The partial regression plot for $\mathbf{z}$ shows its association with the response $\mathbf{y}$ after adjusting for the variables in $\mathbf{X_0}$. This plot is obtained by plotting residuals $\mathbf{r}_y$ from the LS regression of $\mathbf{y}$ on $\mathbf{X}_0$ versus residuals $\mathbf{r}_z$ from the LS regression of $\mathbf{z}$ on $\mathbf{X}_0$.

a) Show that $\mathbf{r}_y = (\mathbf{I} - \mathbf{H}_0)\mathbf{y}$ and $\mathbf{r_z} = (\mathbf{I} - \mathbf{H}_0)\mathbf{z}$, where $\mathbf{H}_0 = \mathbf{X_0}(\mathbf{X_0^T}\mathbf{X_0})^{-\mathbf{1}}\mathbf{X_0^T}$.

Answer: Recall $\hat{\mathbf{y}} = \mathbf{X}_0\hat{\boldsymbol{\beta}} = \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T\mathbf{y}$. Therefore,

$$\begin{aligned}\mathbf{r}_y &= \mathbf{y} - \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T\mathbf{y}\\ &= (\mathbf{I} - \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T)\mathbf{y}\\ &= (\mathbf{I} - \mathbf{H}_0)\mathbf{y}.\end{aligned}$$

Similarly,

$$\mathbf{r}_z = \mathbf{z} - \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T\mathbf{z} = (\mathbf{I} - \mathbf{H}_0)\mathbf{z},$$

which finishes the proof.

b) Under the model assumptions given above, show that $E(\mathbf{r}_y) = \mathbf{r}_z\gamma$, so the expected slope of the line is the same as the coefficient of $\mathbf{z}$ in the full model.

Answer: We compute from a)

$$
\begin{aligned}
E(\mathbf{r}_y) &= E\big((\mathbf{I} - \mathbf{H}_0)\mathbf{y}\big) \\
&= (\mathbf{I} - \mathbf{H}_0)E(\mathbf{y}) && (\mathbf{I} - \mathbf{H}_0 \text{ is constant}) \\
&= (\mathbf{I} - \mathbf{H}_0)(\mathbf{X}_0\boldsymbol{\beta} + \mathbf{z}\gamma + E(\mathbf{e})) \\
&= (\mathbf{I} - \mathbf{H}_0)(\mathbf{X}_0\boldsymbol{\beta} + \mathbf{z}\gamma) && (E(\mathbf{e}) = \mathbf{0}) \\
&= (\mathbf{I} - \mathbf{H}_0)\mathbf{z}\gamma && (\mathbf{X}_0\boldsymbol{\beta} \in \mathrm{col}(\mathbf{X}_0)) \\
&= \mathbf{r}_z\gamma, && (\text{from a}))
\end{aligned}
$$

where $\mathrm{col}(\mathbf{X}_0)$ denotes the column space of $\mathbf{X}_0$ and we use the fact that projection matrix of the space, to which a vector is orthogonal, i.e., $\mathbf{I} - \mathbf{H}_0$, projects this vector into a zero vector.

c) The conditional expectation model in b) has the form of simple linear regression through the origin (no intercept). Show that fitting this "model" by LS regression gives estimated slope:

$$\hat{\gamma} = \frac{\mathbf{z}^{\mathbf{T}}(\mathbf{I} - \mathbf{H_0})\mathbf{y}}{\mathbf{z}^{\mathbf{T}}(\mathbf{I} - \mathbf{H_0})\mathbf{z}} = \frac{\sum_{i=1}^{n}(z_i - \hat{z}_i)y_i}{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}$$

where $\hat{\mathbf{z}} = \mathbf{H}_0\mathbf{z}$ and y_i, z_i and $\hat{z}_i$ are the ith components of $\mathbf{y}$, $\mathbf{z}$ and $\hat{\mathbf{z}}$ respectively.

Answer: Recall the formula of OLS estimates for the model of regression from origin from a) in Problem 2 in HW1. Here the covariates is $\mathbf{r}_z$ and the response is $\mathbf{r}_y$. Thus, we can plug in those quantities into the formula to obtain

$$
\begin{aligned}
\hat{\gamma} = \frac{\mathbf{r}_z^T\mathbf{r}_y}{\mathbf{r}_z^T\mathbf{r}_z} &= \frac{\mathbf{z}^T(\mathbf{I} - \mathbf{H}_0)^T(\mathbf{I} - \mathbf{H}_0)\mathbf{y}}{\mathbf{z}^T(\mathbf{I} - \mathbf{H}_0)^T(\mathbf{I} - \mathbf{H}_0)\mathbf{z}} \\
&= \frac{\mathbf{z}^T(\mathbf{I} - \mathbf{H}_0)\mathbf{y}}{\hat{\mathbf{z}}^T(\mathbf{I} - \mathbf{H}_0)^T(\mathbf{I} - \mathbf{H}_0)\mathbf{z}} \\
&= \frac{\sum_{i=1}^{n}(z_i - \hat{z}_i)y_i}{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2},
\end{aligned}
$$

where we applied the fact that for a projection matrix $\mathbf{P}$, $\mathbf{P} = \mathbf{P}^T$ and $\mathbf{PP} = \mathbf{P}$.

d) Using c) and the model assumptions, show $E(\hat{\gamma}) = \gamma$ and

$$var(\hat{\gamma}) = \frac{\sigma^2}{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}$$

Answer: Since **z** is deterministic, from c), we compute

$$\begin{aligned} E(\hat{\gamma}) = \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)E(\mathbf{y})}{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}} &= \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)(\mathbf{X}_0\boldsymbol{\beta}+\mathbf{z}\boldsymbol{\gamma})}{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}} \\ &= \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}}{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}} \cdot \gamma \\ &= \gamma, \end{aligned}$$

and

$$\begin{aligned} var(\hat{\gamma}) = \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)var(\mathbf{y})(\mathbf{I}-\mathbf{H}_0)\mathbf{z}}{(\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z})^2} &= \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{I}(\mathbf{I}-\mathbf{H}_0)\mathbf{z}}{(\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z})^2} \cdot \sigma^2 \\ &= \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}}{(\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z})^2} \cdot \sigma^2 \\ &= \frac{\sigma^2}{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (z_i - \hat{z}_i)^2}, \end{aligned}$$

which finishes the proof.

e) Let R_z^2 denote the multiple R-square statistic for the regression of **z** on the variables in $\mathbf{X}_0$. It can be shown that

$$R_z^2 = 1 - \frac{\sum_{i=1}^n (z_i - \hat{z}_i)^2}{\sum_{i=1}^n (z_i - \bar{z}_i)^2}$$

where $\bar{z} = n^{-1} \sum_{i=1}^n z_i$. Using this fact, show that

$$var(\hat{\gamma}) = VIF_z * \frac{\sigma^2}{\sum_{i=1}^n (z_i - \bar{z}_i)^2}$$

where VIF_z is the "variance inflation factor" given by

$$VIF_z = \frac{1}{1 - R_z^2}$$

Answer: From d), we compute,

$$\begin{aligned} var(\hat{\gamma}) &= \frac{\sigma^2}{\sum_{i=1}^n (z_i - \hat{z}_i)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (z_i - \bar{z})^2} \cdot \frac{\sum_{i=1}^n (z_i - \bar{z})^2}{\sum_{i=1}^n (z_i - \hat{z}_i)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (z_i - \bar{z})^2} \cdot \frac{1}{1 - R_z^2} \\ &= VIF_z \times \frac{\sigma^2}{\sum_{i=1}^n (z_i - \bar{z})^2}, \end{aligned}$$

which finishes the proof.