5. The next question uses data on the demand for alcohol. The variables are: `Q` - quantity demanded of alcohol, and `P` - the price of alcohol. The sample size is $n = 100$. A least squares model is estimated, and the results are shown below.

```
summary(lm(Q ~ P, data = demand))


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.3225     1.8594   7.165 1.46e-10 ***
P            -0.5938     0.1767  -3.360  0.00111 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.705 on 98 degrees of freedom
Multiple R-squared:  0.1033,    Adjusted R-squared:  0.09416
F-statistic: 11.29 on 1 and 98 DF,  p-value: 0.001111
```

a) Write down the population model (econometric model) that has been estimated. What are some factors that might be in ϵ?

b) How much will alcohol consumption decrease if a tax raises the price by \$1?

c) How much alcohol will be consumed if the price is set at \$10?

d) One of the data points is $P = 10$, $Q = 8.03$. What is the LS residual for this data point?

e) Suppose that it is a widely held belief that people drink 0.5 less when price increases by \$1. Calculate the t-test statistic for this hypothesis test. Decide whether you "reject" or "fail to reject" the null hypothesis.

f) Is the price variable "significant"?

g) Construct a 95% confidence interval around b_1.

h) Interpret the model's R^2.

1. What is the interpretation of the β in the multiple regression model?
2. What is the problem with R-square (R^2), and how does adjusted-R-square ($\bar{R}^2$) fix the problem?
3. Explain omitted variable bias (OVB), using the *fireplaces* and *house price* data (or any other example). What kinds of variables shouldn't be omitted from a model? What happens if these kinds of variables *are* omitted?
4. What is the dummy variable trap, and why is it a problem for LS estimation?

5. This question is about differences-in-differences (DiD). 27 US states decide to adopt universal health care (treatment group), while 23 states do not (control group). The yearly number of doctor visits per person is measured in each state, 1 year before and 1 year after the adoption of universal health care. During this time, there was another global pandemic. We want to estimate the *causal* effect of universal **health care** on the average number of yearly per-person **doctor visits**.

Table 1: Average number of doctor visits for treatment and control groups, before and after the adoption of universal health care.

	$\texttt{time} = 0$	$\texttt{time} = 1$
no universal health care $\texttt{treatment.group} = 0$	2.78	3.28
universal health care $\texttt{treatment.group} = 1$	2.64	3.24

a) What is the DiD estimator for the effect of universal health care on doctor visits?

b) What assumption needs to be made for the DiD estimator to be valid? Comment on the feasibility of this assumption for this example.

c) Explain how the DiD estimator could be obtained by estimating a model using least squares. Write this population model down, and explain which β represents the DiD estimator.

6. This question estimates a polynomial model using the video game data from assignment 2. `Sales` is how much the game sold (in millions) and `Score` is the rating of the game (most games score between 5 and 10).

```
mod <- lm(Sales ~ Score + I(Score ^ 2), data=mydata)
summary(mod)
```



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.70448    0.58114   8.095 7.21e-16 ***
Score       -1.79812    0.17616 -10.207  < 2e-16 ***
I(Score^2)   0.17396    0.01311  13.265  < 2e-16 ***
```

a) Do you think that `Score` has a linear or a non-linear effect on `Sales`? Why?

b) What is the estimated effect of a game scoring a higher rating on its sales? (Don't consider scores less than 5).

7. A **log**-lin model for wages is estimated, which includes **interaction terms**, and uses the CPS data (sample size $n = 534$):

```
mod.unrestricted <- lm(log(wage) ~ education + experience + gender
                       + gender*education + gender*experience, data = cps)
summary(mod.unrestricted)
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.311024   0.188009   1.654 0.098659 .
education             0.111749   0.012641   8.840  < 2e-16 ***
experience            0.008896   0.002435   3.654 0.000284 ***
gendermale            0.401835   0.243155   1.653 0.099009 .
education:gendermale -0.021452   0.016265  -1.319 0.187776
experience:gendermale 0.007449   0.003373   2.209 0.027629 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.449 on 528 degrees of freedom
Multiple R-squared:  0.2831,    Adjusted R-squared:  0.2763
F-statistic:  41.7 on 5 and 528 DF,  p-value: < 2.2e-16
```

a) What is the estimated effect of education on wage, for men and for women?

b) Test the null hypothesis that the interaction terms are not needed (that the effect of education and experience **jointly** do not depend on gender). Use the following **restricted** model:

```
mod.restricted <- lm(log(wage) ~ education + experience + gender, data = cps)
summary(mod.restricted)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.418982   0.122865   3.410 0.000699 ***
education   0.098006   0.008009  12.237  < 2e-16 ***
experience  0.012668   0.001697   7.466 3.42e-13 ***
gendermale  0.255947   0.039409   6.495 1.92e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4523 on 530 degrees of freedom
Multiple R-squared:  0.2696,    Adjusted R-squared:  0.2655
F-statistic: 65.22 on 3 and 530 DF,  p-value: < 2.2e-16
```