Question 1

(a)

Since the variable `approve` takes the value 1 if mortgage is approved and 0 if it is not, we expect the coefficient on `white` to be positive, as we expect the probability of approval to be higher for white people.

(b)

We run a simple OLS with `approve` as the dependent variable and `white` as the explanatory variable. The OLS results is shown below:

```
##
## Call:
## lm(formula = approve ~ white, data = loanapp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90839  0.09161  0.09161  0.09161  0.29221
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.70779    0.01824   38.81   <2e-16 ***
## white        0.20060    0.01984   10.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3201 on 1987 degrees of freedom
## Multiple R-squared:  0.04893,    Adjusted R-squared:  0.04845
## F-statistic: 102.2 on 1 and 1987 DF,  p-value: < 2.2e-16
```

As the coefficient of `white` is positive, we conclude that white people have 20.06% higher chances of mortgage approval compared to those who are not white, holding all else constant. The coefficient of the slope is also positive. However, the interpretation of the slope is not logical because `approve` is a binary variable.

To test if it is statistically significant, we conduct a t-test with the following hypotheses:
$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

According to the result, the p-value for the slope coefficient is practically 0. Therefore, we reject the null hypothesis and conclude that the slope coefficient is practically large, and statistically significant.

(c)

We add the other variables listed and run the OLS again, with the output as follow:

```
##
## Call:
## glm(formula = approve ~ white + hrat + obrat + loanprc + unem +
##     male + married + dep + sch + cosign + chist + pubrec + mortlat1 +
##     mortlat2 + vr, data = loanapp)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.06482   0.00781   0.06387   0.13673   0.71105
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.936731   0.052735  17.763  < 2e-16 ***
## white        0.128820   0.019732   6.529 8.44e-11 ***
## hrat         0.001833   0.001263   1.451   0.1469
## obrat       -0.005432   0.001102  -4.930 8.92e-07 ***
## loanprc     -0.147300   0.037516  -3.926 8.92e-05 ***
## unem        -0.007299   0.003198  -2.282   0.0226 *
## male        -0.004144   0.018864  -0.220   0.8261
## married      0.045824   0.016308   2.810   0.0050 **
## dep         -0.006827   0.006701  -1.019   0.3084
## sch          0.001753   0.016650   0.105   0.9162
## cosign       0.009772   0.041139   0.238   0.8123
## chist        0.133027   0.019263   6.906 6.72e-12 ***
## pubrec      -0.241927   0.028227  -8.571  < 2e-16 ***
## mortlat1    -0.057251   0.050012  -1.145   0.2525
## mortlat2    -0.113723   0.066984  -1.698   0.0897 .
## vr          -0.031441   0.014031  -2.241   0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.09124989)
##
##     Null deviance: 213.79  on 1970  degrees of freedom
## Residual deviance: 178.39  on 1955  degrees of freedom
##   (18 observations deleted due to missingness)
## AIC: 892.51
##
## Number of Fisher Scoring iterations: 2
```

The coefficient on `white` dropped to 0.1288 from 0.2006 since the introduction of the control variables essentially meant that the some of the impact of `white` on `approve` are captured by the control variables.

To test if there is discrimination against non-whites, we conduct the following test:

$H_0 : \beta_{white} = 0$
$H_1 : \beta_{white} \neq 0$

The p-value of white coefficient is still practically 0 albeit larger than before. We reject the null and conclude that there exists statistically significant evidence to conclude discrimination against non-whites. Being white has 12.88% higher chances of mortgage approval compared to those who are not white, holding all else constant.

(d)

We run a probit with only `white` as the explanatory variable. The output is as follow:

```
##
## Call:
## glm(formula = approve ~ white, family = binomial(link = "probit"),
##     data = loanapp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1864   0.4384   0.4384   0.4384   0.8314
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.54695    0.07544   7.251 4.15e-13 ***
## white        0.78395    0.08671   9.041  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1480.7  on 1988  degrees of freedom
## Residual deviance: 1401.8  on 1987  degrees of freedom
## AIC: 1405.8
##
## Number of Fisher Scoring iterations: 4
```

For being white, the probability of loan approval slight decreases from 90.84% drop to 90.82%. On the contrary, the probability of loan approval increases from 70.78% to 70.88% for non-whites by using probit model.

To determine if probit model is significant or not, we conduct a similar t-test as before on the slope coefficient. The p-value is still practically 0, so we reject the null and conclude that probit model is significant. Unexpectedly, there is indeed statistically strong evidence to suggest discrimination against non-whites.

(e)

We add the same control variables as (c) and fit another probit.

```
##
## Call:
## glm(formula = approve ~ white + hrat + obrat + loanprc + unem +
##     male + married + dep + sch + cosign + chist + pubrec + mortlat1 +
##     mortlat2 + vr, family = binomial(link = "probit"), data = loanapp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1018   0.2369   0.3430   0.4884   1.9897
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.062330   0.316308   6.520 7.03e-11 ***
## white        0.520254   0.096866   5.371 7.84e-08 ***
## hrat         0.007876   0.007026   1.121  0.26227
## obrat       -0.027693   0.006147  -4.505 6.64e-06 ***
## loanprc     -1.011956   0.240265  -4.212 2.53e-05 ***
## unem        -0.036685   0.017680  -2.075  0.03799 *
## male        -0.037000   0.109883  -0.337  0.73632
## married      0.265745   0.094724   2.805  0.00502 **
## dep         -0.049575   0.039065  -1.269  0.20443
## sch          0.014648   0.095415   0.154  0.87799
## cosign       0.086064   0.240886   0.357  0.72088
## chist        0.585278   0.095602   6.122 9.24e-10 ***
## pubrec      -0.778742   0.126984  -6.133 8.65e-10 ***
## mortlat1    -0.187628   0.257164  -0.730  0.46563
## mortlat2    -0.494354   0.325883  -1.517  0.12927
## vr          -0.201062   0.081478  -2.468  0.01360 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1476.0  on 1970  degrees of freedom
## Residual deviance: 1200.5  on 1955  degrees of freedom
##   (18 observations deleted due to missingness)
## AIC: 1232.5
##
## Number of Fisher Scoring iterations: 5
```

To figure out if there is discrimination against non-whites, we conduct a hypothesis testing as follow:

$H_0 : \beta_{white} = 0$
$H_1 : \beta_{white} \neq 0$

The p-value of white coefficient is again, practically 0 up till the 8th decimal place, so we reject the null and conclude that there is discrimination against non-whites. Being white has higher chances of mortgage approval compared to those who are not white, holding all else constant. However, since the model is non-linear, we cannot calculate the difference in estimation between the two models as they are different at each combination of explanatory variables.

(f)

We run a logit with the same control variables as (e). The output is as follow:

```
##
## Call:
## glm(formula = approve ~ white + hrat + obrat + loanprc + unem +
##     male + married + dep + sch + cosign + chist + pubrec + mortlat1 +
##     mortlat2 + vr, family = binomial(link = "logit"), data = loanapp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9549   0.2545   0.3458   0.4768   2.0827
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.80171    0.59467   6.393 1.63e-10 ***
## white        0.93776    0.17290   5.424 5.84e-08 ***
## hrat         0.01326    0.01288   1.030  0.30313
## obrat       -0.05303    0.01128  -4.702 2.58e-06 ***
## loanprc     -1.90495    0.46041  -4.138 3.51e-05 ***
## unem        -0.06658    0.03281  -2.029  0.04242 *
## male        -0.06639    0.20642  -0.322  0.74776
## married      0.50328    0.17799   2.828  0.00469 **
## dep         -0.09073    0.07333  -1.237  0.21598
## sch          0.04123    0.17840   0.231  0.81723
## cosign       0.13206    0.44608   0.296  0.76720
## chist        1.06658    0.17121   6.230 4.67e-10 ***
## pubrec      -1.34067    0.21736  -6.168 6.92e-10 ***
## mortlat1    -0.30988    0.46351  -0.669  0.50378
## mortlat2    -0.89468    0.56857  -1.574  0.11559
## vr          -0.34983    0.15372  -2.276  0.02286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1476  on 1970  degrees of freedom
## Residual deviance: 1201  on 1955  degrees of freedom
##   (18 observations deleted due to missingness)
## AIC: 1233
##
## Number of Fisher Scoring iterations: 5
```

Comparing the coefficient on white to probit estimate,`white` is still statistically significant and positive. However, we are not able to directly compare the impact of `white` on `approve` with only the coefficient due to the different specifications of logit and probit.

(g)

To compare the size of the discrimination effect between probit and logit, we construct three test groups.

The first group consists of dummy variables that are all zero and non-dummy variables taking the mean value. The second group has dummy variables equal to 1 and non-dummy variables taking the mean value. All variables other than `white` in the third group is set to its mean. Then, holding all other variables constant, we calculate the discrimination effect by setting `white` to 1 and 0 for each case, finding the predicted probabilities for all test cases in both the probit (part e) and logit models (part f), and then finding the difference in predicted probabilities for each test case.

For the "Zeroes" test case, Being whites has 16.58% chance to get loan approval than non-whites by using probit model (65.55% for non-whites and 82.13% for whites). As for the logit model, the probability will increase 17.61% for being white (64.942% for non-whites and 82.55% for whites).

For the "Ones" test case the difference between whites and non-whites increases to 20.45% and 22.93% for probit and logit models respectively.

Lastly, for the "Means" test case, the result is totally contrary. Being whites has 10.6% higher probability than non-white by using probit model while the difference is only 9.7% for logit model.

In brief, the size of the discrimination effect is larger for logit model when we assume all dummy variables are either 0 or 1 and logit model has smaller size of the discrimination effect by adopting mean for all variables. The reason behind this result can be explained by non-linearity.

Fit with Probit.

white	case	probability
0	Zeroes	0.6554800
1	Zeroes	0.8213212
0	Ones	0.3643359

white	case	probability
1	Ones	0.5688162
0	Means	0.8170743
1	Means	0.9230845

Fit with Logit

white	case	probability
0	Zeroes	0.6493866
1	Zeroes	0.8255063
0	Ones	0.3538988
1	Ones	0.5831745
0	Means	0.8280339
1	Means	0.9249905

Discrimination effect

Case	Probit	Logit
Zeroes	0.1658412	0.1761197
Ones	0.2044803	0.2292757
Means	0.1060102	0.0969566

Question 2

We fit a logit model as follow:

```
##
## Call:
## glm(formula = happymar ~ church + female + educ, family = binomial(link = "logit"),
##     data = marriage)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6519  -0.5434   0.2483   0.7683   2.2158
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.1586     3.2849  -2.484  0.01300 *
## church        2.9075     0.9206   3.158  0.00159 **
## female        2.3945     0.8772   2.730  0.00634 **
## educ          0.5267     0.2651   1.986  0.04699 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 79.763  on 60  degrees of freedom
## Residual deviance: 49.268  on 57  degrees of freedom
## AIC: 57.268
##
## Number of Fisher Scoring iterations: 5
```

(a)

$H_0 : \beta_{church} = 0$
$H_1 : \beta_{church} \neq 0$

The coefficient of `church` is 2.9075 with p-value 0.00159. As the p-value is less than 1%, we can reject the null hypothesis and conclude that coefficient of `church` is significant at 1%.

$H_0 : \beta_{female} = 0$
$H_1 : \beta_{female} \neq 0$

The coefficient of `female` is 2.3945 with p-value 0.00634. As the p-value is less than 1%, we can reject the null hypothesis and conclude that coefficient of `female` is significant at 1%.

$H_0 : \beta_{educ} = 0$
$H_1 : \beta_{educ} \neq 0$

The coefficient of `educ` is 0.5267 with p-value 0.04699. As the p-value is less than 5% but more than 1%, we can only reject the null hypothesis and conclude that coefficient of `female` is significant at 5%.

Therefore, the determinants of happiness are `church`, `female` and `educ`, and the signs of these three variables appear positive, indicating that all three of `church`, `female` and `educ` positively affect the probability of marital happiness, `happymar`.

(b)

Best to create a new dataframe with all the variables needed. The columns `church`, `female`, `educ` are inputs, `log-odds` are estimates from the model, and `odds` is calculated from `log-odds`. Likewise, `P(Happy)` is derived from `odds`.

church	female	educ	log.odds	odds	P.Happy.
0	0	8	-3.945068	0.0193499	0.0189826
1	0	8	-1.037530	0.3543287	0.2616268
0	1	16	2.662934	14.3382955	0.9348037
1	1	16	5.570472	262.5579262	0.9962058

Thus, the above table summarizes the results of log odds, odds and probability of marital happiness for the following individuals:
(1) A male with 8 years of education who does not go to church regularly
(2) A male with 8 years of education who goes to church regularly
(3) A female with 16 years of education who does not go to church regularly
(4) A female with 16 years of education who goes to church regularly

(c)

To calculate McFadden's Pseudo R2, we need to first fit the model with no explanatory variables.

```
##
## Call:
## glm(formula = happymar ~ 1, family = binomial(link = "logit"),
##     data = marriage)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4282  -1.4282   0.9458   0.9458   0.9458
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5725     0.2666   2.147   0.0318 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 79.763  on 60  degrees of freedom
## Residual deviance: 79.763  on 60  degrees of freedom
## AIC: 81.763
##
## Number of Fisher Scoring iterations: 4
```

Then, we calculate McFadden's Pseudo-R by applying the equation, which is defined as $1 - \frac{log(Lfull)}{log(Lnull)}$.

```
PseudoR2 = 1 - logLik(marriage.log)/logLik(marriage.null)
PseudoR2
```

```
## 'log Lik.' 0.3823251 (df=4)
```

The result shows that Pseudo-R2 is 0.3823251, which is greater than the benchmark of 0.2~0.3. Pseudo-R2 represents that approximately 38% of variation in `happymar` is explained by `church`, `female` and `educ`. It appears that the full model with predictors, `church`, `female` and `educ`, provides reasonable predictions of `happymar`.

(d)

Run the updated logit with an additional interaction variable `cheducx`. The output is as follow:

```
##
## Call:
## glm(formula = happymar ~ church + female + educ + cheducx, family = binomial(link = "
##     data = marriage)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7177  -0.5504   0.2409   0.7207   2.1396
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.2654     3.5290  -1.775  0.07583 .
## church        3.1795     1.0190   3.120  0.00181 **
## female        2.3095     0.8724   2.647  0.00812 **
## educ          0.3712     0.2866   1.295  0.19520
## cheducx       0.5493     0.6495   0.846  0.39768
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 79.763  on 60  degrees of freedom
## Residual deviance: 48.439  on 56  degrees of freedom
## AIC: 58.439
##
## Number of Fisher Scoring iterations: 6
```

$H_0 : \beta_{cheducx} = 0$
$H_1 : \beta_{cheducx} \neq 0$

The coefficient of `cheducx` is 0.5493 with p-value of 0.19520. As the p-value is much greater than 5%, we cannot reject the null hypothesis and thus conclude that coefficient of `cheducx` is insignificant at 5%. That is, there's no significant interaction effect between `church` and `educ`.

The insignificance is possibly caused by the fact that `cheducx` is closely related to `educ` as it is an unavoidable and dependent interaction. Therefore, to test whether `cheducx` and `educ` are independent, it is essential to perform a likelihood ratio test between model without interaction, `cheducx`, and model with interaction `cheducx`. The result is as follow:

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
57	49.26757	NA	NA	NA

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
56	48.43854	1	0.8290271	0.362554

It is not significant since the p-value of chi-squared distribution with df=1 is 0.3626, which is greater than 5% of significance level. The two models, with and without `cheducx`, are not statistically different. That is, `cheducx` does not improve the model in a statistically significant way. This result is consistent with the p-value on the coefficient on `cheducx` from the logit.

The test-statistic is computed by the equation of $-2 * log(L_{full}) + 2 * log(L_{reduced})$. The test statistic is essentially a magnified log-ratio of the likelihoods of the reduced model and the full model, and since it is sufficiently small (can't be rejected against a null of 0), we conclude that `cheducx` does not improve the predictability of the model.

```
-2*logLik(marriage.log) + 2*logLik(marriage.d)
```

```
## 'log Lik.' 0.8290271 (df=4)
```

Question 3

(a)

Given that $\beta_{YX} = \frac{Cov(Y,X)}{Var(X)}$, $\beta_{YZ} = \frac{Cov(Y,Z)}{Var(Z)}$, and $\beta_{XZ} = \frac{Cov(X,Z)}{Var(Z)}$, we can derive a consistent estimator of β_1 from the "reduced form" as follows:

To construct a consistent estimator of β_1, we need to show that an exogenous change in x_i is associated with a change in Y_i so that the effect on Y of an exogenous unit change in X is β_1.

From the regression result, we can calculate the regression lines $X_i = \beta_{0XZ} + \beta_{XZ} \times Z_i + e_{XZ}$
$Y_i = \beta_{0XZ} + \beta_{YZ} \times Z_i + e_{YZ}$

Where e_{XZ} and e_{YZ} are the error terms, and Z_i is uncorrelated with e_{XZ} and e_{YZ} by definition.

Next, rearrange the X equation to solve for Z:

$Z_i = -\frac{\beta_{0XZ}}{\beta_{XZ}} + \frac{1}{\beta_{XZ}} X_i - \frac{1}{\beta_{XZ}} e_{XZ}$

Then substitute Z into the Y equation and collect terms

Y_i

$= \beta_{0YZ} + \beta_{YZ} Zi + e_{YZ}$

$= \beta_{0YZ} + \beta_{YZ}(-\frac{\beta_{0XZ}}{\beta_{XZ}} + \frac{1}{\beta_{XZ}} X_i - \frac{1}{\beta_{XZ}} e_{XZ}) + e_{YZ}$

$= (\beta_{0YZ} - \beta_{0XZ}\frac{\beta_{YZ}}{\beta_{XZ}}) + \frac{\beta_{YZ}}{\beta_{XZ}} X_i + (e_{YZ} - \frac{\beta_{YZ}}{\beta_{XZ}} e_{XZ})$

$= \beta_0 + \beta_1 X_i + e_{YX}$

Thus, $\hat{\beta_0} = \beta_{0YZ} - \beta_{0XZ}\frac{\beta_{YZ}}{\beta_{XZ}}$, $\hat{\beta_1} = \frac{\beta_{YZ}}{\beta_{XZ}}$, and $e\hat{Y}_X = e_{YZ} - \frac{\beta_{YZ}}{\beta_{XZ}} e_{XZ}$

From above derivation, we know in the regression model $\beta_1 = \frac{Cov(Y,Z)}{Cov(X,Z)}$. When the sample size is greater than 100, $\hat{\beta}_1 = \frac{\beta_{YZ}}{\beta_{XZ}} = \frac{Cov\hat{(}Y,Z)}{Cov(\hat{X},Z)}$, or the fraction of sample covariances. We know that sample covariances are consistent estimators of population covariance, therefore the ratio of them is also consistent. Additionally, the interpretation is that an exogenous change in X_i of β_{XZ} units is associated with a change in Y_i of βYZ units when the sample size is large. The effect on Y of an exogenous unit change in X is $\hat{\beta}_1 = \frac{\beta_{YZ}}{\beta_{XZ}}$.

Therefore, $\hat{\beta}_1$ is a consistent estimator of β_1.

(b)

$\hat{\beta}_1$ is a coefficient estimator of TSLS = SYZ/SXZ = $\frac{\sum_{i=1}^{n} Y_i(Zi-\bar{Z})}{\sum_{i=1}^{n} X_i(Zi-\bar{Z})}$ Then, substitute in Yi= β_0+ $\beta_1 Xi$+ $e.YX$, multiply through n^1/2 and simplify:

We get Var($\hat{\beta}_1$) = $\frac{Var[(Zi-\bar{Z})*e_{YX}]}{n[Cov(X_i,Z_i)]^2}$, the variance of TSLS coefficient estimator whereas the variance of OLS estimator is, Var($\hat{\beta}_1$) = $\frac{Var(e.YX)}{\sum_{i=1}^{n}(Xi-\bar{X})}$.

Therefore, as shown in part a, it appears that the $\hat{\beta}_1$ is a consistent estimator of standard IV estimator, β_1 with the same expected value. As for the variance of the calculated estimator, it is biased against the variance of OLS estimator as proved in part B. The main reason is because we isolate the part of X that is uncorrelated with the error term in the first stage of TSLS calculation to solve the endogenous.It is also noticeable that the COV(X,Z) should be much greater than 0 so that application of central limit theorem is applicable and the statistical inferences are reasonably following the approximate normal distribution.

Question 4

(a)

We know that labour demand is dependent on real wage while labour supply is simply a straight line. Then, the general equations are as follow: $L_D(w) = \alpha_0 + \alpha_1 w + e_d$
$L_S = \beta_0 + e_s$
where w is real wage, and e_d and e_s are the error terms for labour demand and labour supply respectively.

(b)

$\hat{\beta}_0$ is simply the mean of labour in the sample, since it is not dependent on any other explanatory variables. Therefore, the equation can be estimated as: $L_S = \bar{L}$

(c)

Want to show that $cov(w, e_d) \neq 0$. That is, we want to show that w can be written as a function of e_d.

Since labour market clears, $L_D(w) = L_S$ holds. That is,

$\alpha_0 + \alpha_1 w + e_d = \beta_0 + e_s$

Rearranging the above and isolating for w gives:

$w = -\frac{\alpha_0}{\alpha_1} + \frac{\beta_0}{\alpha_1} + \frac{e_s}{\alpha_1} - \frac{1}{\alpha_1} e_d$

Let $-\frac{\alpha_0}{\alpha_1} + \frac{\beta_0}{\alpha_1} + \frac{e_s}{\alpha_1} = \gamma$, then:

$w(e_d) = \gamma - \frac{1}{\alpha_1} e_d$

w can be written as a function of e_d, which is a random variable. Therefore, it follows that $cov(w, e_d) \neq 0$.

(d)

We should not estimate the labour demand equation using OLS if real wage is indeed endogenous. Instead we should use fit an 2SLS to remove the endogeity issue. The endogenous explanatory variable would violate the exogeneity assumptions of the Gauss-Markov theorem of OLS. Continued use of OLS models in this case would lead to biased estimators and poor predictions as the expected value from the model would deviate from the true expected value of our parameter (labour quantity). An OLS model in this case should not be used for any policy recommendations.

(e)

We would need to select an instrument that shifts labour supply but not labour demand. Labour participation rate of females is an obvious choice. It definitely increases labour supply (though total effect is countered by other variables) but there's no reason to think that it shifts demand in any way (i.e. employers need to hire similar number of workers regardless of how many women are in the workforce), therefore it is exogenous and not correlated with e_d

It is also relevant, as higher labour participation rate of females lead to lower real wage on its own given the influx of supply in the labour market.

Question 5

Importing data

(a)

The coefficient should be positive for `democrat`, `vote_1` and `lispend`. Democrats won the election, so the votes and two-party votes should increase if the candidate belongs to democrat party.

The higher the previous vote share, the higher two-party votes and votes. It is an indication that people have been willing to vote for this candidate. It also deters the entrance of a good inter-party challenger.

Same for the incumbent spending because the more incumbent spendings mean that the candidate has more money for campaign, resulting in the higher votes and two-party votes.

As for `hq` and `lcspend`,the coefficient for them should be negative.

The higher the challenger quality, the lower two-party votes and votes. There are few possible reasons behind it. Firstly,a good challenger could run as an independent, therefore splitting the votes of the two main parties. Also, a tough primary battle could damage the incumbent in the press for the general election. Lastly, people may think that there must be a lot of people vote for this god challenger, so it doesn't matter if they go to vote or not, which results in the lower two-party votes and votes.

For the higher challenger spending,it indicate that the challenger has more money on advertisement, which can be viewed as a threat for incumbent. Therefore,it will reduce the two-party votes and votes.

(b)

The OLS result is as follow:

```
##
## Call:
## glm(formula = vote ~ democrat + hq + vote_1 + lispend + lcspend,
##     data = spending)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -14.6681   -3.3813    0.2593    3.3448   13.4733
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.68000    5.10888   7.375 2.55e-12 ***
## democrat     4.53049    0.64642   7.009 2.34e-11 ***
## hq          -2.53573    0.88394  -2.869  0.00448 **
## vote_1       0.51738    0.04446  11.637  < 2e-16 ***
## lispend      0.16837    0.57204   0.294  0.76875
## lcspend     -1.85352    0.20492  -9.045  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 23.21858)
##
##     Null deviance: 19131.4  on 249  degrees of freedom
## Residual deviance:  5665.3  on 244  degrees of freedom
## AIC: 1503.6
##
## Number of Fisher Scoring iterations: 2
```

After we run the OLS regression and compare the results with our hypothesis, we found that it is same as what we predict. However, `lispend` is not significant because its p-value is

bigger than 5%.

(c)

The first reason could be location of incumbent's office. For example, a candidate from an urban New York riding typically spends more than a candidate from rural Montana just because it is more expensive to run ads in New York than Montana. Therefore, location of incumbent's office will affect its spending and the vote.

Another reason is traveling time and expense. If incumbent spend more time on traveling and advertising, it is more likely to increase the spending and vote rate.

(d)

The variable incumbent spending in the previous election cycle (`lispend1`) is likely correlated with current incumbent spending and it typically not related to current election's challenger quality. Therefore, we can fix the endogenity issue by isolating the part of incumbent spending that is uncorrelated with the error terms using `lispend`, provided that we still include controls for geographical variables.

(e)

We run 2 separate OLS and see what happens. First stage has `lispend` as the response variable with `democrat`, `hq`, `vote_1` and `lcspend` as control. `lispend1` is proposed as an instrument. The output is as follow:

```
##
## Call:
## glm(formula = lispend ~ democrat + hq + vote_1 + lcspend + lispend1,
##     data = spending)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.99501  -0.22250   0.00744   0.20842   1.19791
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.949187   0.447739   2.120    0.035 *
## democrat    0.046531   0.050871   0.915    0.361
## hq          0.094923   0.069063   1.374    0.171
## vote_1      0.006214   0.003771   1.648    0.101
## lcspend     0.096765   0.015483   6.250 1.82e-09 ***
## lispend1    0.704766   0.044307  15.907  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1427628)
```

```
##
##     Null deviance: 101.906  on 249  degrees of freedom
## Residual deviance:  34.834  on 244  degrees of freedom
## AIC: 230.75
##
## Number of Fisher Scoring iterations: 2
```

Now, we run the second stage OLS with `vote` as the response and the same controls in the first stage as independent variables along with the exogenous `lispend_pred` from the first stage. The output is as follow:

```
##
## Call:
## glm(formula = vote ~ democrat + hq + vote_1 + lispend_pred +
##     lcspend, data = spending)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13.4362   -3.1197    0.1869    3.5090   13.1049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.40570    6.25755   4.060 6.61e-05 ***
## democrat      4.57484    0.63833   7.167 9.07e-12 ***
## hq           -2.95596    0.88184  -3.352  0.00093 ***
## vote_1        0.55640    0.04544  12.243  < 2e-16 ***
## lispend_pred  2.00637    0.79153   2.535  0.01188 *
## lcspend      -2.02968    0.20918  -9.703  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 22.63089)
##
##     Null deviance: 19131.4  on 249  degrees of freedom
## Residual deviance:  5521.9  on 244  degrees of freedom
## AIC: 1497.2
##
## Number of Fisher Scoring iterations: 2
```

(f)

Now, we run 2SLS directly in R without separating the 2 stages. The output is as follow:

```
##
## Call:
## ivreg(formula = vote ~ democrat + hq + vote_1 + lispend + lcspend |
```

```
##     democrat + hq + vote_1 + lispend1 + lcspend, data = spending)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.042  -3.677   0.108   3.742  13.195
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.40570    6.47098   3.926 0.000112 ***
## democrat     4.57484    0.66010   6.931 3.72e-11 ***
## hq          -2.95596    0.91192  -3.241 0.001355 **
## vote_1       0.55640    0.04699  11.840  < 2e-16 ***
## lispend      2.00637    0.81853   2.451 0.014941 *
## lcspend     -2.02968    0.21631  -9.383  < 2e-16 ***
##
## Diagnostic tests:
##                  df1 df2 statistic  p-value
## Weak instruments   1 244    253.02  < 2e-16 ***
## Wu-Hausman         1 243     11.15 0.000972 ***
## Sargan             0  NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.919 on 244 degrees of freedom
## Multiple R-Squared: 0.6913,  Adjusted R-squared: 0.685
## Wald test: 112.5 on 5 and 244 DF,  p-value: < 2.2e-16
```

Like part (e), `lispend` is significant with the same slope coefficient of 2.00637. However, it has a slightly larger standard error than (g) because OLS standard errors assume unconditional homoskedasticity while TSLS is conditional on Z. In other words, we have to use inconsistent standard errors for TSLS. Either of these results are not unexpected given the model specifications.

(g)

For a instrument to be valid, it needs to be relevant and exogenous.

To test if instrument is relevant, we compute the first-stage F-statistic from the first stage of OLS.

$H_0 : \beta_{lispend1} = 0$ (weak relevance)
$H_0 : \beta_{lispend1} \neq 0$ (strong relevance)

The test output is as follow.

Res.Df	Df	F	Pr(>F)
245	NA	NA	NA

Res.Df	Df	F	Pr(>F)
244	1	253.0181	0

We reject the null and conclude that the instrument is strong because F is bigger than 10. We could've used the output from the "weak instruments" test in (f) and obtained the same result. It is relevant.

Then, we can test for the exogeneity of `lispend1` by computing the correlation between `lispend1` and the residuals from the OLS as follow:

```
cor.test(spending$lispend1, spending.b$residuals, method=c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  spending$lispend1 and spending.b$residuals
## t = 1.9521, df = 248, p-value = 0.05205
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.001063624  0.243372110
## sample estimates:
##       cor
## 0.1230196
```

Based on the p-value of 0.05205, we conclude that `lispend1` is not correlated with the error terms of the OLS, therefore it is exogenous.

Based on the result of TSLS, `lispend1` is a valid instrument for the endogenous `lispend`.

(h)

Re refit a TSLS, but this time with 2 instruments `lispend1` and `hq`.

```
##
## Call:
## ivreg(formula = vote ~ democrat + lcspend + vote_1 + lispend |
##     democrat + vote_1 + lcspend + lispend1 + hq, data = spending)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5874  -3.5881   0.3471   3.4741  13.6631
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.68543    6.44419   4.451 1.30e-05 ***
## democrat     4.70059    0.66440   7.075 1.56e-11 ***
```

```
## lcspend     -2.12962    0.21587  -9.865  < 2e-16 ***
## vote_1       0.55002    0.04734  11.618  < 2e-16 ***
## lispend      1.46188    0.80773   1.810   0.0715 .
##
## Diagnostic tests:
##                   df1 df2 statistic p-value
## Weak instruments   2 244   132.071 < 2e-16 ***
## Wu-Hausman         1 244     7.975 0.00513 **
## Sargan             1  NA    10.547 0.00116 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.96 on 245 degrees of freedom
## Multiple R-Squared: 0.6849,  Adjusted R-squared: 0.6798
## Wald test: 135.7 on 4 and 245 DF,  p-value: < 2.2e-16
```

`lispend` is endogenous in this model, since the Wu-Hausman test can be rejected.

To test if instrument is relevant, The "Weak instruments" p-value of near 0 and F-test statistic of 145.668 suggests that `lscpend` is indeed relevant as an instrument for `lispend`. We reject the null and conclude that the instruments are strong because F is still bigger than 10.

Then, we conduct the second-stage J-test to test for exogeneity and see if we have over-identified.

H_0: Both lispend1 and hq are uncorrelated with the error terms (instruments are exogenous)
H_1: lispend1 or/and hq are correlated with the error terms (at least one instrument is endogenous)

The result of the J-test is given in the output as "Sargan" with the test-statistic of 10.547 and p-value of 0.00116. At least one of the instruments is endogenous and should be removed from the model. Comparing this result to (g), we can conclude that we can obtain a good TSLS model with only `lispend1` as the instrument. Whether or not `hq` is a good instrument on its own needs to be re-fitted.

We could also compute the p-value by hand as per the Anderson-Rubin J-test and obtain the same result.

Therefore, using both `lispend1` and `hq` together results in over-identification and it makes at least one of them endogenous. We can definitely use `lispend` alone as an instrument, but we have to conduct further tests on `hq` to conclude the same thing.

```
spending.h2 = glm(residuals(spending.h)~democrat+vote_1+
                    lcspend+lispend1+hq, data=spending)
spendingLHTest = linearHypothesis(spending.h2,
                                  c("lispend1 = 0", "hq = 0"), test="Chisq")
pchisq(spendingLHTest[2, 3], df = 1, lower.tail = FALSE)
```

```
## [1] 0.001044614
```

(i)

We would suggest any geographical variables as instrument for both incumbent spending and challenger spending, such as the location of office, the state that candidate runs for campaign. As mentioned in part(c), these geographical factors are correlated with incumbent spending(lispend), but it won't affect other variables in the regression. They are also relevant with the traveling time and expense. For example,if the incumbent's office is at a place with well-developed transit system, incumbent is more likely to travel around, which increases the spending. Therefore, location can overcome both reasons in part(c) and it is a valid instrument.

Same reason can be applied to challenger spending. The challenger's office location, and traveling time and expense could be the reasons that challenger spending is endogenous.