Problem Set 1

Pencil and Paper Questions

1. (a) i.

$$TSS = ESS + RSS$$

$$TSS = \sum_{i=1}^{n} y_i^2$$

$$ESS = \sum_{i=1}^{n} \widehat{y_i^2}$$

$$RSS = \sum_{i=1}^{n} e_i^2$$

$$\boldsymbol{TSS} = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y} + \hat{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left((\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)\right)^2$$

$$= \sum_{i=1}^{n} \left((\hat{y}_i - \bar{y}) + \hat{e}_i\right)^2 = \sum_{i=1}^{n} \left((\hat{y}_i - \bar{y})^2 + 2\hat{e}_i(\hat{y}_i - \bar{y}) + \hat{e}_i^{\,2}\right) =$$

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} \hat{e}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} \hat{e}_i^{\,2}$$

$$= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{e}_i^{\,2} + 2\sum_{i=1}^{n} \hat{e}_i\left(\widehat{\beta_0} + \widehat{\beta_1} x_{i1} + (\dots) + \widehat{\beta_k} x_{ik} - \bar{y}\right) =$$

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{e}_i^{\,2} + 2\left(\widehat{\beta_0} - \bar{y}\right)\sum_{i=1}^{n} \hat{e}_i + 2\widehat{\beta_1}\sum_{i=1}^{n} \hat{e}_i\, x_{i1} + (\dots) + 2\widehat{\beta_k}\sum_{i=1}^{n} \hat{e}_i\, x_{ik} =$$

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{e}_i^{\,2} = \sum_{i=1}^{n} \widehat{y_i^2} + \sum_{i=1}^{n} \hat{e}_i^{\,2} = \boldsymbol{ESS} + \boldsymbol{RSS}$$

(a) ii.

$$\boldsymbol{R^2} = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n} \widehat{y_i^2}}{\sum_{i=1}^{n} y_i^2} = \frac{ESS}{ESS + RSS} = \frac{\sum_{i=1}^{n} \widehat{y_i^2}}{\sum_{i=1}^{n} \widehat{y_i^2} + \sum_{i=1}^{n} e_i^2} = \frac{\sum_{i=1}^{n} \widehat{y_i^2} + \sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} \widehat{y_i^2} + \sum_{i=1}^{n} e_i^2} =$$

$$\frac{\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} = \boldsymbol{1 - \frac{RSS}{TSS}}$$

(b)

$$R^2 = \rho_{y,\hat{y}}^2$$

$$\rho_{y,\hat{y}} = \frac{\sigma_{y,\hat{y}}^2}{\sqrt{\sigma_y^2 * \sigma_{\hat{y}}^2}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 * \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 * \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}} =$$

$$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 * \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(y_i\hat{y}_i - \hat{y}_i^{\,2} + \hat{y}_i^{\,2} - \bar{y}\hat{y}_i - \bar{y}y_i + \hat{y}_i\bar{y} - \bar{y}\hat{y}_i + \bar{y}^2)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 * \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}} =$$

$$\frac{\sum_{i=1}^{n}((y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 * \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 * \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}} =$$

$$\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \sqrt{\frac{ESS}{TSS}} \rightarrow \boldsymbol{\rho^2_{y,\hat{y}}} = \frac{\boldsymbol{ESS}}{\boldsymbol{TSS}} = \boldsymbol{R^2}$$

As shown above the R^2 is the squarred correlation between y and ,$\hat{y}$ for regressions. The correlation is in a range between -1 and 1 whereas the R^2 is between 0 and 1. The correlation explains the deviation form a curve of points (y) and how appropriate the estimations is of it. The R^2 explains the variation in the estimated $\hat{y}_i$ and its true value y_i and the correlation does exactly the same with an other point of view.

(c)

Recap: The $\hat{\beta}$ and $\tilde{\beta}$ are calculated as follows ($\tilde{X} = 2X$):

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$\tilde{\boldsymbol{\beta}} = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'y = \left((2X)'(2X)\right)^{-1}(2X)'y = (2X'2X)^{-1}2X'y = (4X'X)^{-1}2X'y$$
$$= 4^{-1}(X'X)^{-1}2X'y = 4^{-1} \cdot 2(X'X)^{-1}X'y = \frac{2}{4}(X'X)^{-1}X'y = \frac{\mathbf{1}}{\mathbf{2}}\hat{\boldsymbol{\beta}}$$

```
Call:
lm(formula = data1$consumption ~ data1$income)

Residuals:
   Min     1Q Median     3Q    Max
-32507  -4310  -2025   1953 137831

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.800e+03  1.328e+02   43.69   <2e-16 ***
data1$income 2.667e-01  6.275e-03   42.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8605 on 6369 degrees of freedom
Multiple R-squared:  0.2209,    Adjusted R-squared:  0.2208
F-statistic:  1806 on 1 and 6369 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = data1$consumption ~ income_2)

Residuals:
   Min     1Q Median     3Q    Max
-32507  -4310  -2025   1953 137831

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.800e+03  1.328e+02   43.69   <2e-16 ***
income_2    1.333e-01  3.138e-03   42.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8605 on 6369 degrees of freedom
Multiple R-squared:  0.2209,    Adjusted R-squared:  0.2208
F-statistic:  1806 on 1 and 6369 DF,  p-value: < 2.2e-16
```

The R^2 does not change if $\tilde{X} = 2X$(red). It is just a linear transformation which does not change the relation. The only impact is the calculated $\frac{1}{2}\hat{\beta}$ what can be seen in green. It does not change the underlying data, otherwise there would be a lot of leeway to manipulate the X and therefore receive higher R^2. That would mean for example if you X measured in tonns you could convert it into kilograms ($\tilde{X} = \frac{1}{1000}X$) and you would get a different R^2, this makes intuitively no sense.

(d)

Adding explanatory variables to a model always increases R^2. Namely, if you have 100 observations and run a regression of y_i on $x_i'\beta$ with 100 elements in x_i' you will perfectly predict your dependent variable.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

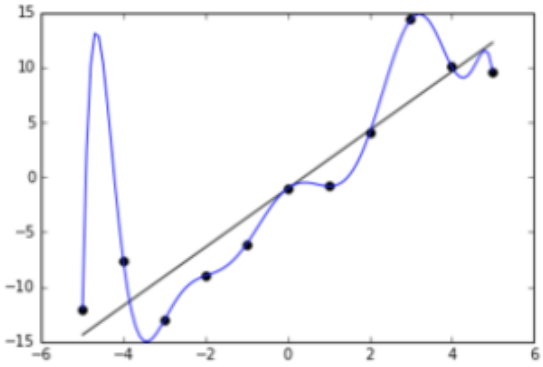
That $R^2 = 1$:

$$RSS = \sum_{i=1}^{n} e_i^2 = 0$$

With increasing explanatory variables, you can better describe your datapoints. That means, if you have for example 11 datapoints and you set up a regression model with 11 explanatory variables, each of the datapoint can be exactly described (shown in the graphic below[1]). So, the

[1] Source of the graphic: “top01d.clrm.hypothesis.testing.topost.INTERIM.SENT” Slide: 20

$\hat{y} = y$ and the error term is zero and therefore the R^2 must become 1. But this leads to the danger of overfitting the model (see 1.e)).



Normal regression:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - x_i\hat{\beta} - \bar{e}\right)^2}{\sum_{i=1}^{n} y_i^2}$$

So, if we add an extra regressor x_i the $\hat{\beta}$ will increase and therefore the RSS is getting smaller as a direct consequence of this. In other words, the precision of the prediction is increasing until the error term is equal 0.

(e)

There are three dangers of R^2:

- Different types of data imply different „typical“ values of R^2:
  - Cross-section data often have low R^2
  - Time-series data often have high R^2
- Different functional forms for y_i can change R^2
- Adding explanatory variables to a model always increase R^2 (see 1.d))

Further it is only an in-sample measurement that means R^2only measures the precision within the sample. A better way to measure the precision is trough out of sample prediction for example with Cross-Validation.

Empirical Application

1. (a)

The following regression model measures the effect of income on total consumption:

$$C_i = \beta_1 + \beta_2 Y_i + \epsilon_i$$

The estimation of this model leads to the following results:

	Dependent variable:
	consumption
income	0.267***
	(0.006)
Constant	5,800.441***
	(132.779)
Observations	6,371
R2	0.221
Adjusted R2	0.221
Residual Std. Error	8,604.836 (df = 6369)
F Statistic	1,805.951*** (df = 1; 6369)
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficient of income $\hat{\beta}_2$ is 0.267. This means that income has a positive effect on total consumption, i.e. an increase in income by 1 USD causes an increase in consumption by 0.267 USD. The coefficient as well as the constant of regression are significant on 99% confidence level. The R^2 has a value of 0.221.

(b)

The following multiple regression model measures the effect of both income and family size on total consumption:

$$C_i = \beta_1 + \beta_2 Y_i + \beta_3 N_i + \epsilon_i$$

Estimating this model gives the following results:

	Dependent variable:
	consumption
income	0.254***
	(0.006)
fam_size	625.431***
	(75.731)
Constant	4,429.220***
	(212.165)
Observations	6,371
R2	0.229
Adjusted R2	0.229
Residual Std. Error	8,559.794 (df = 6368)
F Statistic	946.605*** (df = 2; 6368)
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficient on income is 0.254, which means that an increase in income by 1 USD increases consumption by 0.254 USD when controlling for size of the family. So, the effect stays nearly the same as in model 2.a). The coefficient on family size is 625.431. This means that an increase in family size by one person increases consumption by 625.431 USD when controlling for income. The impact compared to the income on the consumption is heavily bigger, that seems logic because a bigger family (increasing size) has normally more people, and therefore the consumption will increase automatically.
The coefficient as well as the constant are highly significant and the adjusted R^2 is 0.229, what means a slicly increase.

(c)
The following multiple regression model measures the effect of income, family size, and owning a home on total consumption:

$$C_i = \beta_1 + \beta_2 Y_i + \beta_3 N_i + \beta_4 House_i + \epsilon_i$$

This model leads to the following estimation results:

	Dependent variable:
	consumption
income	0.254***
	(0.006)
fam_size	625.445***
	(75.726)
house	1,395.781
	(1,021.504)
Constant	4,413.947***
	(212.445)
Observations	6,371
R2	0.229
Adjusted R2	0.229
Residual Std. Error	8,559.212 (df = 6367)
F Statistic	631.779*** (df = 3; 6367)
Note:	*p<0.1; **p<0.05; ***p<0.01

The value of the coefficient on income is again 0.254 and therefore equal to the coefficient in the previous estimation. Therefore, owning a home does not have any influence on the effect of income on consumption when holding family size constant. The impact of family size stays also relatively constant compared to the question b). The new variable house has e very big coefficient but is not significant on any common statistical level, that is why we would exclude the coefficient house. This also can be seen in a stable adjusted R^2.

(d)
The coefficients on income and their interpretation change whenever the model changes. In the first model, only income is considered as s dependent variable. In the second and the third model, the variables family size and owning a home are added as control variables.
Therefore, the coefficient on income provides estimates of the effect of income on consumption when holding family size and owning a home constant. In the third model, for example, the

coefficient on income suggest that for families of the same size and with a house, an increase in income by 1 UDS increases consumption by 0.254 USD.
We should go for the second model because the coefficient does not change a lot and the adjusted R^2 increases. Furthermore, in the first model, where family size and owning a home are ignored, the effect on income on consumption is overestimated. The regressor "house" should not be taken into account because it is not significant on any common statistical level.

The Gender Wage Gap

2. (a)

The "educ" coefficient increases by adding the independent variable "age" to the regression model. In the first regression we force one variable to explain the wage differences (famous "omitted variable bias", OVB). In the second regression, we have two explanatory variables, which are both significant. Also, in Model 1, we underestimate the explanatory potential of our data. In Model 2, we see considerably increased R^2 and adjusted R^2. A possible interpretation is to see age as a somehow confounding variable, which influences the dependent variable as well as the other so-called "independent" variable (educ) (OVB phenomena). In total that means for the regression two that ceteris paribus an increase of education by 1 year increases the wage by 7.170 thousand $ on average.

$Wage_i = \beta_1 + \beta_2 educ_i$

	Dependent variable:
	wage
educ	7.074***
	(0.028)
Constant	-58.037***
	(0.446)
Observations	561,076
R2	0.104
Adjusted R2	0.104
Residual Std. Error	58.751 (df = 561074)
F Statistic	65,026.500*** (df = 1; 561074)
Note:	*p<0.1; **p<0.05; ***p<0.01

$Wage_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i$

	Dependent variable:
	wage
educ	7.170***
	(0.027)
age	1.461***
	(0.009)
Constant	-115.916***
	(0.554)
Observations	561,076
R2	0.147
Adjusted R2	0.147
Residual Std. Error	57.312 (df = 561073)
F Statistic	48,428.710*** (df = 2; 561073)
Note:	*p<0.1; **p<0.05; ***p<0.01

(b)

The new model is $Lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i$. This model is a Log-Level model. Ceteris paribus, if we increase educ by one unit, we expect the wages to increase by 12% one average. And, also ceteris paribus, if we increase age by one unit, we expect the wages to increase by 2,5% on average. But why is this? For that, we compare the expected values of lw for values of age that differ by Δage_i from age_i. Ceteris paribus, the expected value for lw for a given age is the result of our estimated model ($Lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i$). When age is $age_i + \Delta age_i$, then the expected value is $(Lw_i + \Delta Lw_i) = \beta_1 + \beta_2 educ_i + \beta_3 (age_i + \Delta age_i)$. Thus, the difference is given by $(Lw_i + \Delta Lw_i) - Lw_i = \beta_1 + \beta_2 educ_i + \beta_3 (age_i + \Delta age_i) - (\beta_1 + \beta_2 educ_i + \beta_3 age_i) = \beta_3 \Delta age_i$. According to the holy wisdom of mathematics[2], the following rule can be applied when $\beta_3 \Delta age_i$ is small: $(Lw_i + \Delta Lw_i) - Lw_i \cong (\Delta Lw_i)/Lw_i$. So, $\frac{\Delta Lw_i}{Lw_i} = \beta_3 \Delta age_i$ and if $\Delta age_i = 1$, then $\frac{\Delta Lw_i}{Lw_i} = \beta_3$. So, we can translate this into percentages when the independent variable increases by one unit. We cannot do that in the previous model because there was no logarithm of the dependent variable *wages*.

[2] Stock and Watson: Introduction to Econometrics, updated third global edition, p.318.

So, which of the models ($Wage_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i$ and $Lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i$) is the better model or makes more sense? Unfortunately, we must not look at the R^2-measurements as our basis of decision-making as the dependent variable is not the same in both models. So, we should use our experts' knowledge on the topic, which tells us that a percentage change is better understandable and more common when discussing and comparing wages. So, we go for that one.

	Dependent variable:
	lw
educ	0.120***
	(0.0004)
age	0.025***
	(0.0001)
Constant	0.713***
	(0.009)
Observations	561,076
R2	0.158
Adjusted R2	0.158
Residual Std. Error	0.920 (df = 561073)
F Statistic	52,686.040*** (df = 2; 561073)
Note:	*p<0.1; **p<0.05; ***p<0.01

(c)

The coefficient of female is -0.443. Ceteris paribus, a woman earns 44,3% less than a man on average. In this model, this factor is economically highly important because wages seem to rely heavily on the respective gender type. And the factor is the biggest in our model. The gender type does immensely influence the monetary output and therefore is economically significant in this setting. However, almost certainly there exist other variables which can explain lower wages for women (OVB).

	Dependent variable:
	lw
educ	0.126***
	(0.0004)
age	0.024***
	(0.0001)
female	-0.443***
	(0.002)
Constant	0.815***
	(0.009)
Observations	561,076
R2	0.206
Adjusted R2	0.206
Residual Std. Error	0.894 (df = 561072)
F Statistic	48,385.030*** (df = 3; 561072)
Note:	*p<0.1; **p<0.05; ***p<0.01

In the following we provide a one-sided and two-sided test, because in our opinion both are relevant. It depends on the underlying assumption due to the wage of female. In (**1**) we test two-sided if the difference is equal 0 and (**2**) we assume ex-ante that the effect will/should be negative and therefore we provide a one-sided test.

(1) We want to find out if we can reject H_0. Either large positive or large negative values of $\hat{\beta}_k$ – and thus large positive or negative values of t would cause us to reject H_0.[3] We

[3] Lecture slides.

do not care if β_4 is bigger or smaller than zero. What we want to know is if there exists a non-multicollinearity – we want to reject $\beta_4 = 0$.

```
t test of coefficients:

              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  0.81453555  0.01013008   80.408 < 2.2e-16 ***
educ         0.12647519  0.00051668  244.787 < 2.2e-16 ***
age          0.02419001  0.00013601  177.858 < 2.2e-16 ***
female      -0.44266137  0.00244112 -181.335 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 1: H_0: $\beta_{female} = 0$, H_1: $\beta_{female} \neq 0$

Step 2: $N = 561'076$

$K = 3$ → dof: $561'073$

$$s^2 = \frac{1}{N-K}\sum_{i=1}^{n} e_i^2$$

$$t - \text{Value: } \frac{(\hat{\beta}_{female} - \beta_{female})}{stderr(\hat{\beta}_{female})} = \frac{-0.44266137 - 0}{0.00244112} = (-181.335)$$

Step 3: $\alpha = 5\%$

$P(t_{(561'073)} \leq \overline{t_{0.975}})) = 0.975$

→ $\overline{t_{0.975}} = 1.96$

$|-181.335| > 1.96$

→ *Therefore, we reject H_0.*

The calculated P Value is 0.00000000000000022 and our $\alpha = 0.05$, so we already reject H_0.

That means the hypothesis $\beta_{female} = 0$ is wrong.

(2) The appropriate test here is a one-sided hypothesis test as we expect β_4 to be negative ex ante.

R provides the following data. The computed t-test is clear- its highly significant.

```
t test of coefficients:

              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  0.81453555  0.01013008   80.408 < 2.2e-16 ***
educ         0.12647519  0.00051668  244.787 < 2.2e-16 ***
age          0.02419001  0.00013601  177.858 < 2.2e-16 ***
female      -0.44266137  0.00244112 -181.335 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By hand:

Step 1: H_0: $\beta_{female} > 0$, H_1: $\beta_{female} < 0$

Step 2: $N = 561'076$

$K = 3$

$$s^2 = \frac{1}{N-K}\sum_{i=1}^{n} e_i^2$$

R provides the t- value for female. As we don't trust R we also calculate it "by hand" as this is proposed in the task:

$$t - Value: \frac{(\hat{\beta}_{female} - \beta_{female})}{stderr(\hat{\beta}_{female})} = \frac{-0.44266137 - 0}{0.00244112} = -181.335$$

As expected, the lack of trust in R was not justified. The t value is the same.

Step 3:

We select a common level of significance: $\alpha = 5\%$. The inverse t-calculator implemented in R tells us that for N-K=561'073 degrees of freedom the critical value for a one-tailed test is 1.645.

$P(t_{(561'073)} \leq \overline{t_{0.95}})) = 0.95$
$\rightarrow \overline{t_{0.955}} = 1.645$
$|-181.335| > 1.645$

Therefore, we reject H_0.

(d)

To obtain β_4, we have to estimate three different regression models.

Model 1: $lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i + e_{1i}$
Then, we have to get the residuals from this model (call it e_1)

Model 2: $female_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i + e_{2i}$
Also, get the residuals (call it e_2)

Model 3: $e_{1i} = \alpha_1 + \beta_4 e_{2i} + e_{3i}, E(\alpha_1) = 0$
What we see now dramatically strengthens our trust in the theory: β_4 from Model 3 has the same value as in the previous model $lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i + \beta_4 female_i$. Partitioned regression works!

```
================================================
                        Dependent variable:
                    ---------------------------
                              resid4model1
------------------------------------------------
resid4model2                    -0.443***
                                (0.002)

Constant                         0.000
                                (0.001)

------------------------------------------------
Observations                    561,076
R2                               0.056
Adjusted R2                      0.056
Residual Std. Error       0.894 (df = 561074)
F Statistic         33,493.160*** (df = 1; 561074)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
> |
t test of coefficients:

               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  0.81453555  0.01013008   80.408 < 2.2e-16 ***
educ         0.12647519  0.00051668  244.787 < 2.2e-16 ***
age          0.02419001  0.00013601  177.858 < 2.2e-16 ***
female      -0.44266137  0.00244112 -181.335 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(e)

And here is the prove why the theory works as seen in exercise (d).

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y$$

The residual maker:

$$e = y - X\hat{\beta} = y - X(X'X)^{-1}X'y = (I_N - X(X'X)^{-1}X')y$$

where: $M_x = (I_N - X(X'X)^{-1}X')$ (indempotent matrix)

$$e = M_x y$$

M_x are square and symmetric and indempotent.

$$y = X\beta + \epsilon = X_k\beta_k + X_{-k}\beta_{-k} + \epsilon$$

$$\hat{\beta} = \begin{bmatrix}\beta_k \\ \beta_{-k}\end{bmatrix} = \begin{bmatrix}X_k'X_k & X_k'X_{-k} \\ X_{-k}'X_k & X_{-k}'X_{-k}\end{bmatrix}^{-1}\begin{bmatrix}X_k'y \\ X_{-k}'y\end{bmatrix} = \begin{bmatrix}(X_k'M_{-k}X_k)^{-1}X_k'M_{-k}y \\ (X_{-k}'M_kX_{-k})^{-1}X_{-k}'M_ky\end{bmatrix}$$

$$\widehat{\beta_k} = (X_k'M_{-k}X_k)^{-1}X_k'M_{-k}y = (X_k'M_{-k}M_{-k}X_k)^{-1}X_k'M_{-k}M_{-k}y$$
$$= (X_k'M_{-k}'M_{-k}X_k)^{-1}X_k'M_{-k}'M_{-k}y = \left(\boldsymbol{X_k^{*\prime}X_k^*}\right)^{-1}\boldsymbol{X_k^{*\prime}y^*}$$

Because: $X_k^* = M_{-k}X_k$ and $y^* = M_{-k}y$

The k[th] coefficient in a multiple OLS regression is equivalent to the coefficient in a simple OLS regression of the residual from a regression of y on all the other regressors ($y^* \equiv M_{-k}y$) on the residual from a regression of X_k on all the other regressions ($X_k^* \equiv M_{-k}X_k$).

$$X_k^* = M_{-k}X_k$$
$$\boldsymbol{y^* = M_{-k}y}$$
$$X_{-k} = \iota = \begin{bmatrix}1 \\ \vdots \\ 1\end{bmatrix}$$
$$M_\iota = (I_N - X(X'X)^{-1}X') = I_N - \iota(\iota'\iota)^{-1}\iota'$$
$$y^* = M_\iota y = (I_N - \iota(\iota'\iota)^{-1}\iota')y \xrightarrow{\iota'\iota = \sum_{i=1}^{N} 1 = N} (I_N - \iota\left(\frac{1}{N}\right)\iota')y = y - \bar{y}$$

And similarly: $X_k^* = M_\iota y X_k = X - \bar{X}$

$$\widehat{\beta_1} = \bar{y} - \bar{X}_{-1}'\widehat{\beta_{-1}}$$

Calucaltion with R

Calculated with R gives exactly the constant term of the estimated model:

```
=====
0.815
-----
```

(f)

The output of the regression is hereby shown:

	Dependent variable:
	lw
educ	0.120***
	(0.001)
age	0.028***
	(0.0002)
female	-0.342***
	(0.018)
educ:female	0.017***
	(0.001)
age:female	-0.009***
	(0.0003)
Constant	0.759***
	(0.011)
Observations	561,076
R2	0.208
Adjusted R2	0.208
Residual Std. Error	0.893 (df = 561070)
F Statistic	29,436.260*** (df = 5; 561070)
Note:	*p<0.1; **p<0.05; ***p<0.01

Individual Hypothesis:

```
t test of coefficients:

              Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  0.75852555  0.01244459  60.952 < 2.2e-16 ***
educ         0.11961567  0.00063962 187.011 < 2.2e-16 ***
age          0.02840090  0.00017682 160.620 < 2.2e-16 ***
female      -0.34189109  0.02121265 -16.117 < 2.2e-16 ***
educ:female  0.01653551  0.00107602  15.367 < 2.2e-16 ***
age:female  -0.00943893  0.00027603 -34.195 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to R, the t values of female, educ:female and age:female are significant on a two tailed test as the p-values are smaller than the values on the significance code.

Joint Hypothesis:

So, now we test the joint hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$. To do that, it is necessary to implement a F test in R. The output is shown below:

Hypothesis test with standard errors under homoskedasticity

```
Linear hypothesis test

Hypothesis:
female = 0
educ:female = 0
age:female = 0

Model 1: restricted model
Model 2: lw ~ educ + age + female + female * (educ) + female * (age)

  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1 561073 474989
2 561070 446949  3     28040 11733 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the compiled computer test we can reject our hypothesis.

$$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(N-K)} = \frac{(R_U^2 - R_R^2)/q}{(1-R_U^2)/(N-K)}$$

	Dependent variable:
	lw
educ	0.120***
	(0.0004)
age	0.025***
	(0.0001)
Constant	0.713***
	(0.009)
Observations	561,076
R2	0.158
Adjusted R2	0.158
Residual Std. Error	0.920 (df = 561073)
F Statistic	52,686.040*** (df = 2; 561073)
Note:	*p<0.1; **p<0.05; ***p<0.01

According to our compiled R Code, $R_U^2 = 0.208$ and $R_R^2 = 0.158$. q is the number of restrictions/hypotheses. In this case, q equals 3. For N, we have 561'076. For K, we have 5. This applied equals:

$$F = \frac{\frac{0.208 - 0.158}{3}}{\frac{1 - 0.208}{561'076 - 2}} = 11'733.15$$

F is bigger than the critical value the internet provides us with the given N and K. So, we reject H_0 as R suggests. What we also see is that our test by hand equals the F-test result under homoscedasticity

(g)

		Regression
(1)	Model in 2f)	$lw_i = 0.759 + 0.12educ_i + 0.028age_i - 0.342female_i + 0.017female_i * educ_i - 0.009female_i * age_i$
(2)	Model for females only	$lw_i = 0.417 + 0.136educ_i + 0.019age_i$
(3)	Model for males only	$lw_i = 0.759 + 0.12educ_i + 0.028age_i$

Comparison of (3) and (1): The coefficients do not change.
Comparison of (2) and (1): The coefficients change.

Interacting a dummy variable with all the other variables changes the regression model. Whereas the male model stays the same as nothing changes at all, the female model differs from the interacting model. This is because the dummy interaction only has an impact on the regression when the female dummy is true. In our opinion, this can make sense to adapt the model and fit it better.

(h)

Please see the attached RCode to find out how we managed to generate the dummies. Of course, it is technically possible to include all of them into the model. If that makes sense is a separate question. But if we include all of them and the constant term into our regression, we get perfect multicollinearity ("dummy variable trap"). To avoid that, it makes sense to exclude a binary variable. So, let us drop "occupationother" as this is the dummy we do not know exactly what it describes. Nevertheless, we already covered the case of this dummy by including all other dummies. So, it does not only make sense to include all dummies except one- we are somehow obliged to do that.

Statistic	N
age	561,121
incwage	561,121
female	561,121
childrenly	561,121
educ	561,076
occupationbusiness	561,121
occupationhealthcare	561,121
occupationother	561,121
occupationscience	561,121
occupationtechnology	561,121
wage	561,121
lw	561,121

(i)

Including all dummies except the "occupation-other" dummy leads us to the following model:

$$lw_i = 1.083 + 0.106educ_i + 0.023age_i - 0.442female_i + 0.549occupationbusiness_i + 0.396occupationhealthcare_i + 0.393occupationscience_i + 0.468occupationtechnology_i$$

The output is displayed here:

	Dependent variable:
	lw
educ	0.106***
	(0.0004)
age	0.023***
	(0.0001)
female	-0.442***
	(0.002)
occupationbusiness	0.549***
	(0.004)
occupationhealthcare	0.396***
	(0.005)
occupationscience	0.393***
	(0.018)
occupationtechnology	0.468***
	(0.005)
Constant	1.083***
	(0.009)
Observations	561,076
R2	0.240
Adjusted R2	0.240
Residual Std. Error	0.874 (df = 561068)
F Statistic	25,353.590*** (df = 7; 561068)
Note:	*p<0.1; **p<0.05; ***p<0.01

On average you earn 54.9% more working in business than in others. A female who works in business earns ca. 10.7% (54.9% -44.2%) more than a female who is working in other businesses. Nevertheless, a woman still earns 44.2% less than men in the same occupation.

Ok, let us perform a model ($lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i + \beta_4 female_i$) for each occupational subsample.

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	1.09 ***	0.94 ***	1.94 ***	1.85 ***	1.13 ***
	(0.04)	(0.04)	(0.12)	(0.04)	(0.01)
educ	0.12 ***	0.14 ***	0.06 ***	0.09 ***	0.10 ***
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
age	0.03 ***	0.02 ***	0.03 ***	0.02 ***	0.02 ***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
female	-0.36 ***	-0.36 ***	-0.14 ***	-0.23 ***	-0.46 ***
	(0.01)	(0.01)	(0.03)	(0.01)	(0.00)
R^2	0.24	0.23	0.21	0.15	0.16
Adj. R^2	0.24	0.23	0.21	0.15	0.16
Num. obs.	45466	40998	2440	30665	441507
RMSE	0.75	0.75	0.63	0.62	0.91

*** p < 0.001, ** p < 0.01, * p < 0.05

Model 1: occupation-business=1
Model 2: occupation-healthcare=1
Model 3: occupation-science=1
Model 4: occupation-technology=1
Model 5: occupation-other=1

The wage gap between male and female seems to be dependent on the occupation. The strongest effect we see in Model 5: -46.3%, the weakest effect in Model 3: -14%. It is to state, that also the other coefficients and constants change in dependence on the occupation. Just a few remarks: the biggest constant is the science constant- people seem to have rather high wages there. Education has a low influence in science- because science people have a similar level of education- otherwise they would not work in a scientist field.
As the coefficients are negative and statistically highly significant for all subsamples, it can be concluded that the hypothesis H0: $b_{female} = 0$ can be rejected. Thus, there is a gap in wages between male and female, a so-called gender wage gap.

(j) i

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	1.51 ***	0.94 ***	0.99 ***	0.73 ***	2.54 ***
	(0.10)	(0.05)	(0.04)	(0.02)	(0.16)
educ	0.10 ***	0.13 ***	0.12 ***	0.11 ***	0.03 ***
	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)
age	0.02 ***	0.02 ***	0.02 ***	0.02 ***	0.03 ***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
childrenly	-0.02	-0.05	-0.01	-0.07 ***	0.23 **
	(0.06)	(0.03)	(0.02)	(0.01)	(0.08)
R^2	0.13	0.17	0.14	0.11	0.14
Adj. R^2	0.13	0.17	0.14	0.11	0.13
Num. obs.	5549	18689	31298	185638	1025
RMSE	0.65	0.74	0.71	0.95	0.56

*** p < 0.001, ** p < 0.01, * p < 0.05

Model 1: occupation-technology=1
Model 2: occupation-business=1
Model 3: occupation-healthcare=1
Model 4: occupation-other=1
Model 5: occupation-science=1

For workers in technology, R states that the childrenly effect is negative and not significant on any common level. In business the effect is significant on a 10% level (-5.2%), in healthcare the effect is negative but not significant on any common level, in other businesses the effect is highly statistically significant (-7,1%), and in science the childrenly effect is positive on 99%. There, childrenly increases your wage on 23,2%.

	Dependent variable:
	lw
educ	0.104***
	(0.004)
age	0.018***
	(0.001)
childrenly	-0.020
	(0.050)
Constant	1.509***
	(0.091)
Observations	5,549
R2	0.128
Adjusted R2	0.128
Residual Std. Error	0.655 (df = 5545)
F Statistic	272.404*** (df = 3; 5545)
Note:	*p<0.1; **p<0.05; ***p<0.01

t test of coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.50871765	0.10390927	14.5196	<2e-16	***
educ	0.10383415	0.00507403	20.4638	<2e-16	***
age	0.01813396	0.00099285	18.2645	<2e-16	***
childrenly	-0.01997854	0.05759283	-0.3469	0.7287	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$$H_0: \beta_{childrenly} > 0, H_1: \beta_{childrenly} < 0$$

$$t - Value: \frac{(\hat{\beta}_{childrenly} - \beta_{childrenly})}{stderr(\hat{\beta}_{childrenly})} = \frac{-0.01997854 - 0}{0.05759283} = -0.3469$$

So the P-Value is .364678., we cannot reject $H_0: \beta_{childrenly} > 0$.

As we just compare a female dataset the interpretation of the results is very difficult. In businesses where the effect is not significant it does not matter if you had children in the last year- You earn the same. In businesses where the effect is negative your wage is negatively impacted. This would be what we expect because you are not working in this time, and other not-pregnant women work on.

(j) ii

The regression output is shown below:

	Dependent variable:
	lw
educ	0.114***
	(0.001)
age	0.018***
	(0.0002)
childrenly	-0.057***
	(0.009)
occupationbusiness	0.601***
	(0.007)
occupationhealthcare	0.407***
	(0.006)
occupationscience	0.534***
	(0.028)
occupationtechnology	0.639***
	(0.012)
Constant	0.704***
	(0.014)
Observations	242,199
R2	0.194
Adjusted R2	0.194
Residual Std. Error	0.900 (df = 242191)
F Statistic	8,347.067*** (df = 7; 242191)
Note:	*p<0.1; **p<0.05; ***p<0.01

R says it is significant on a 5% level:

```
Linear hypothesis test

Hypothesis:
occupationbusiness - occupationscience = 0

Model 1: restricted model
Model 2: lw ~ educ + age + childrenly + occupationbusiness + occupationhealthcare +
    occupationscience + occupationtechnology

  Res.Df    RSS Df Sum of Sq    F  Pr(>F)
1 242192 196233
2 242191 196228  1    4.3671 5.39 0.02025 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

By hand:

```
t test of coefficients:

                      Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)          0.70391130  0.01677259  41.9680 < 2.2e-16 ***
educ                 0.11429259  0.00086436 132.2279 < 2.2e-16 ***
age                  0.01769793  0.00020826  84.9808 < 2.2e-16 ***
childrenly          -0.05651227  0.01030980  -5.4814 4.224e-08 ***
occupationbusiness   0.60108150  0.00595172 100.9929 < 2.2e-16 ***
occupationhealthcare 0.40681874  0.00485634  83.7707 < 2.2e-16 ***
occupationscience    0.53388243  0.01881771  28.3713 < 2.2e-16 ***
occupationtechnology 0.63865161  0.00923228  69.1759 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

	(Intercept)	educ	age	childrenly	occupationbusiness	occupationhealthcare	occupationscience	occupationtechnology
(Intercept)	0.0002	-0.00001	-0.00000	-0.00001	0.00001	0.00001	0.00003	0.00001
educ	-0.00001	0.00000	0	-0.00000	-0.00000	-0.00000	-0.00000	-0.00000
age	-0.00000	0	0.00000	0.00000	-0.00000	-0.00000	0.00000	-0.00000
childrenly	-0.00001	-0.00000	0.00000	0.0001	-0.00000	-0.00000	0.00000	0.00000
occupationbusiness	0.00001	-0.00000	-0.00000	-0.00000	0.00005	0.00001	0.00001	0.00001
occupationhealthcare	0.00001	-0.00000	-0.00000	-0.00000	0.00001	0.00003	0.00001	0.00001
occupationscience	0.00003	-0.00000	0.00000	0.00000	0.00001	0.00001	0.001	0.00001
occupationtechnology	0.00001	-0.00000	-0.00000	0.00000	0.00001	0.00001	0.00001	0.0002

$$t - Value:$$

$$\frac{(\hat{\beta}_{occupationbusiness} - \beta_{occupationsience} - 0)}{\sqrt{s^2(X'X)^{-1}_{\beta_{business}} + s^2(X'X)^{-1}_{\beta_{science}} - 2s^2(X'X)^{-1}_{\beta_{science},\beta_{business}}}}$$

$$= \frac{0.60108150 - 0.53388243 - 0}{\sqrt{0.900^2 * 0.00005 + 0.900^2 * 0.001 - 2 * 0.900^2 * 0.00001}} = 2.32650$$

$$\sqrt{5.39} = 2.322 \approx 2.32650$$

The result calculated by hand is slightly bigger than the one calculated by R.

(j) iii

They are not the same because we compare female with all dataset. In tendency, the effects are all bigger. So, it is better for women to work in a specific sector than in the "other" occupation. See it displayed below:

Female model

	Dependent variable:
	lw
educ	0.114***
	(0.001)
age	0.018***
	(0.0002)
childrenly	-0.057***
	(0.009)
occupationbusiness	0.601***
	(0.007)
occupationhealthcare	0.407***
	(0.006)
occupationscience	0.534***
	(0.028)
occupationtechnology	0.639***
	(0.012)
Constant	0.704***
	(0.014)
Observations	242,199
R2	0.194
Adjusted R2	0.194
Residual Std. Error	0.900 (df = 242191)
F Statistic	8,347.067*** (df = 7; 242191)
Note:	*p<0.1; **p<0.05; ***p<0.01

All model

	Dependent variable:
	lw
educ	0.106***
	(0.0004)
age	0.023***
	(0.0001)
female	-0.442***
	(0.002)
occupationbusiness	0.549***
	(0.004)
occupationhealthcare	0.396***
	(0.005)
occupationscience	0.393***
	(0.018)
occupationtechnology	0.468***
	(0.005)
Constant	1.083***
	(0.009)
Observations	561,076
R2	0.240
Adjusted R2	0.240
Residual Std. Error	0.874 (df = 561068)
F Statistic	25,353.590*** (df = 7; 561068)
Note:	*p<0.1; **p<0.05; ***p<0.01

(k)

Because of assumption 2 we could provide causal statements. It is a nice assumption for these purposes. As soon we let it fall, we must consider the error term. It is possible, or to express it pessimistically, likely, that in this error term some dependent variables remain which we have not considered yet. One factor we can think of is p.e. the years of work experience. Women often have less working experience than men because it was usual that after giving birth the mother takes some years off. Another explanatory variable could be different social behavior between men and women- which is difficult to measure.

R Code

```
## Pencil and Paper Example: c)
setwd("~/Dropbox/Empirical Methods/Problemset 2/Grunddaten")
data1 <- read.csv("consumption.csv", header=TRUE, sep = ",")
head(data1)
test1 <- lm(data1$consumption~data1$income)
summary(test1)
income_2 <- 2*data1$income
test2 <- lm(data1$consumption~income_2)
summary(test2)

## Empirical Application
## 1. Coefficient Interpretation
#Install Package stargazer
install.packages("stargazer")

#Set Working Directory
setwd("~/Dropbox/Empirical Methods/Problemset 2/Grunddaten")

#Load packages
library(stargazer)
library(sandwich)
library(zoo)
library(lmtest)
#Load data and Attach ce
mydata <-read.table("consumption.csv", header=T, sep=",")
mydata$house
attach(mydata)

#a) Regression a
#Regress consumption on income
reg_a <- lm(consumption ~ income)
summary(reg_a)
stargazer(reg_a, type = "text", out = "reg_a.htm")

#b) Regress consumption on income and familysize
reg_b <- lm(consumption ~ income + fam_size)
summary(reg_b)
stargazer(reg_b, type = "text", out = "reg_b.htm")

#c)
reg_c <- lm(consumption ~ income + fam_size + house)
summary(reg_c)
stargazer(reg_c, type = "text", out = "reg_c.htm")
##Gender wage gap
setwd("C:/Users/Markus/Dropbox/Empirical Methods/Problemset 2/Working Progress/RCode")
library(lmtest)
library(stargazer)
library(dplyr)
```

```
library(car)
library(dummies)
library(sandwich)

data1 = read.csv("sampleUScens2015.csv")  # read csv file

#(a)
data2 = mutate(data1, wage=incwage/1000, lw=log(wage)) #add two variables
head(data2)
model1 <- lm(wage ~ educ, data = data2) #regress wage on education
stargazer(model1, type="text")
model2 <- lm(wage ~ educ + age, data = data2) #regress wage on education and age
stargazer(model2, type="text")


#(b)
model3 <- lm(lw ~ educ + age, data = data2)
stargazer(model3,type="text")

#(c)
model4 <- lm(lw ~ educ + age + female, data = data2)
stargazer(model4,type="text")
#USE F-TEST BECAUSE ITS EASIEST AND MOST GENERIC WAY TO DO THAT STUFF
fitted.model <- model4
R <- rbind(c(0,0,0,1))
r <- 0
ftest <- linearHypothesis(model4, hypothesis.matrix=R, rhs=r, vcov=vcovHC(fitted.model,"HC1"))
ftest

###HERE, PROVIDE US WITH T-test stuff for calculation by hand
model.str.elpct <- model4
model.coeftest <- coeftest(model.str.elpct, vcov=vcovHC(model.str.elpct,"HC1"))
model.coeftest

#CHECK-APPROACH
coeftest(model4,vcov=vcovHC(model4,"HC1"))
hoho <- coeftest(model4,vcov=vcovHC(model4,"HC1"))
betafemale<-hoho[4] #row matrix
betafemale<-model.coeftest[4]

criticalvalueforfemale <- abs(qt(0.05, 561073)) #critical value for a one tailed test on alpha 5% needed for calculation by hand

#does not work:
#If (betafemale<criticalvalueforfemale) print("reject") else print("fail to reject")

#(d) partitioned regression
model4a <- lm(lw ~ educ + age, data = data2)#regression w/o female
resid4model1 <- resid(model4a)
```

```
model4b <- lm(female ~ educ + age, data=data2) #regress female on all x
resid4model2 <- resid(model4b)
model4c <- lm(resid4model1~resid4model2)
stargazer(model4c,type="text")

#(e) proof
betairgendwas <- matrix(coefficients(model4))
betairgendwas1 <- mean(data2$lw)-cbind(mean(data2$educ,na.rm=TRUE),mean(data2$age),
mean(data2$female))%*%betairgendwas[2:4]
stargazer(betairgendwas1,type="text")

#(f)
model5 <- lm(lw ~ educ + age + female + female*(educ) + female*(age), data = data2)
stargazer(model5,type="text")
unrestricted<-summary(model5)$r.squared

#joint hypothesis test
fitted.model <- model5
R <- rbind(c(0,0,0,1,0,0), c(0,0,0,0,1,0), c(0,0,0,0,0,1))
r <- c(0,0,0)
ftest <- linearHypothesis(model5, hypothesis.matrix=R, rhs=r,
vcov=vcovHC(fitted.model,"HC1"))
ftest
ftestwithoutvcov <- linearHypothesis(model5, hypothesis.matrix=R, rhs=r)
ftestwithoutvcov
#for on hand calculations i need r^2 unrestricted and restricted,
#i already have unrestricted r2
wow_model <- lm(lw ~ educ + age, data = data2)
stargazer(wow_model,type="text")
restricted<-summary(wow_model)$r.squared


#F test by hand
fvalue=((unrestricted-restricted)/3)/((1-unrestricted)/(561076-2))
fvalue

fvalue1=((unrestricted1-restricted1)/3)/((1-unrestricted1)/(561076-5))
fvalue1

#ttest for individual hypothesis
coeftest(model5,vcov=vcovHC(model5,"HC1"))


#(g)
model6 <- lm(lw ~ educ + age, data = subset(data2,female==0)) #for males
stargazer(model6,type="text")
model7 <- lm(lw ~ educ + age, data = subset(data2,female==1)) #for females
stargazer(model7,type="text")

#(h)
data99 <- dummy.data.frame(data2[,c(1,2,3,4,5,7,9,10)],sep="") #createdummyvariables
```

```
names(data99) #check
stargazer(data99,type="text")

#(i)
model8 <- lm(lw ~ educ + age + female + occupationbusiness +
             occupationhealthcare + occupationscience +
             occupationtechnology,data = data99) #dont include occupationalother
stargazer(model8,type="text")

#foreachoccupationalsubsample
model1001 <- lm(lw ~ educ + age + female, data = subset(data99,occupationbusiness==1))
stargazer(model1001,type="text") #Case1
model1002 <- lm(lw ~ educ + age + female, data = subset(data99,occupationhealthcare==1))
stargazer(model1002,type="text") #Case2
model1003 <- lm(lw ~ educ + age + female, data = subset(data99,occupationscience==1))
stargazer(model1003,type="text") #Case3
model1004 <- lm(lw ~ educ + age + female, data = subset(data99,occupationtechnology==1))
stargazer(model1004,type="text") #Case4
model1005 <- lm(lw ~ educ + age + female, data = subset(data99,occupationother==1))
stargazer(model1005,type="text") #Case5

stargazer(model1001,model1002,model1003,model1004,model1005, out="name.htm",
type="text")

model1001.se <-sqrt ( diag ( vcovHC ( model1001,"HC1")))
model1002.se <-sqrt ( diag ( vcovHC ( model1002,"HC1")))
model1003.se <-sqrt ( diag ( vcovHC ( model1003,"HC1")))
model1004.se <-sqrt ( diag ( vcovHC ( model1004,"HC1")))
model1005.se <-sqrt ( diag ( vcovHC ( model1005,"HC1")))
library ( texreg )
screenreg ( list ( model1001, model1002, model1003, model1004, model1005) ,
        override.se= list ( model1001.se , model1002.se , model1003.se , model1004.se ,
model1005.se ))

#(j)
femaledata = subset(data99,female==1) #malesaway
names(femaledata)
femaledata2 <- dummy.data.frame(femaledata[,c(1,2,4,5,6,7,8,9,10,11,12)],sep="")
#createdummyvariables
names(femaledata2) #check

#(j)(i)
###technology workers
femalemodel1 <- lm(lw ~ educ + age + childrenly ,
                data = subset(femaledata2,occupationtechnology==1)) #regress
stargazer(femalemodel1, type="text")
coeftest(femalemodel1, vcov = vcovHC(femalemodel1,"HC1"))
#do that linear hypothesis stuff
coeftest(femalemodel1,vcov=vcovHC(femalemodel1,"HC1"))
#trial linearHypothesis(femalemodel1, c("childrenly"))
###business workers
```

```
femalemodel2 <- lm(lw ~ educ + age + childrenly ,
                   data = subset(femaledata2,occupationbusiness==1)) #regress
stargazer(femalemodel2, type="text")
###healthcare workers
femalemodel3 <- lm(lw ~ educ + age + childrenly ,
                   data = subset(femaledata2,occupationhealthcare==1)) #regress
stargazer(femalemodel3, type="text")
###other workers
femalemodel4 <- lm(lw ~ educ + age + childrenly ,
                   data = subset(femaledata2,occupationother==1)) #regress
stargazer(femalemodel4, type="text")
###science workers
femalemodel5 <- lm(lw ~ educ + age + childrenly ,
                   data = subset(femaledata2,occupationscience==1)) #regress
stargazer(femalemodel5, type="text")

femalemodel1.se <-sqrt ( diag ( vcovHC ( femalemodel1,"HC1")))
femalemodel2.se <-sqrt ( diag ( vcovHC ( femalemodel2,"HC1")))
femalemodel3.se <-sqrt ( diag ( vcovHC ( femalemodel3,"HC1")))
femalemodel4.se <-sqrt ( diag ( vcovHC ( femalemodel4,"HC1")))
femalemodel5.se <-sqrt ( diag ( vcovHC ( femalemodel5,"HC1")))
library ( texreg )
screenreg ( list ( femalemodel1, femalemodel2, femalemodel3, femalemodel4, femalemodel5)
,
        override.se= list ( femalemodel1.se , femalemodel2.se , femalemodel3.se ,
femalemodel4.se , femalemodel5.se ))



#(j)(ii)
model9 <- lm(lw ~ educ + age + childrenly + occupationbusiness +
               occupationhealthcare + occupationscience +
               occupationtechnology,data = femaledata)
stargazer(model9,type="text")

#hypothesistest
mod <- lm(lw ~ educ + age + childrenly + occupationbusiness +
            occupationhealthcare + occupationscience +
            occupationtechnology,data = femaledata)
linearHypothesis(mod,c("occupationbusiness                                    =
occupationscience"),VCOV=VCOVHC(mod,"HC1"))

#ttest
coeftest(model9,vcov=vcovHC(model9,"HC1"))
cov(femaledata$occupationbusiness,femaledata$occupationscience)
a <- vcov(model9, complete="true")
stargazer(a,out="covmatrix.htm",type="text")

#(j)(iii)
#see above
#(k)
```