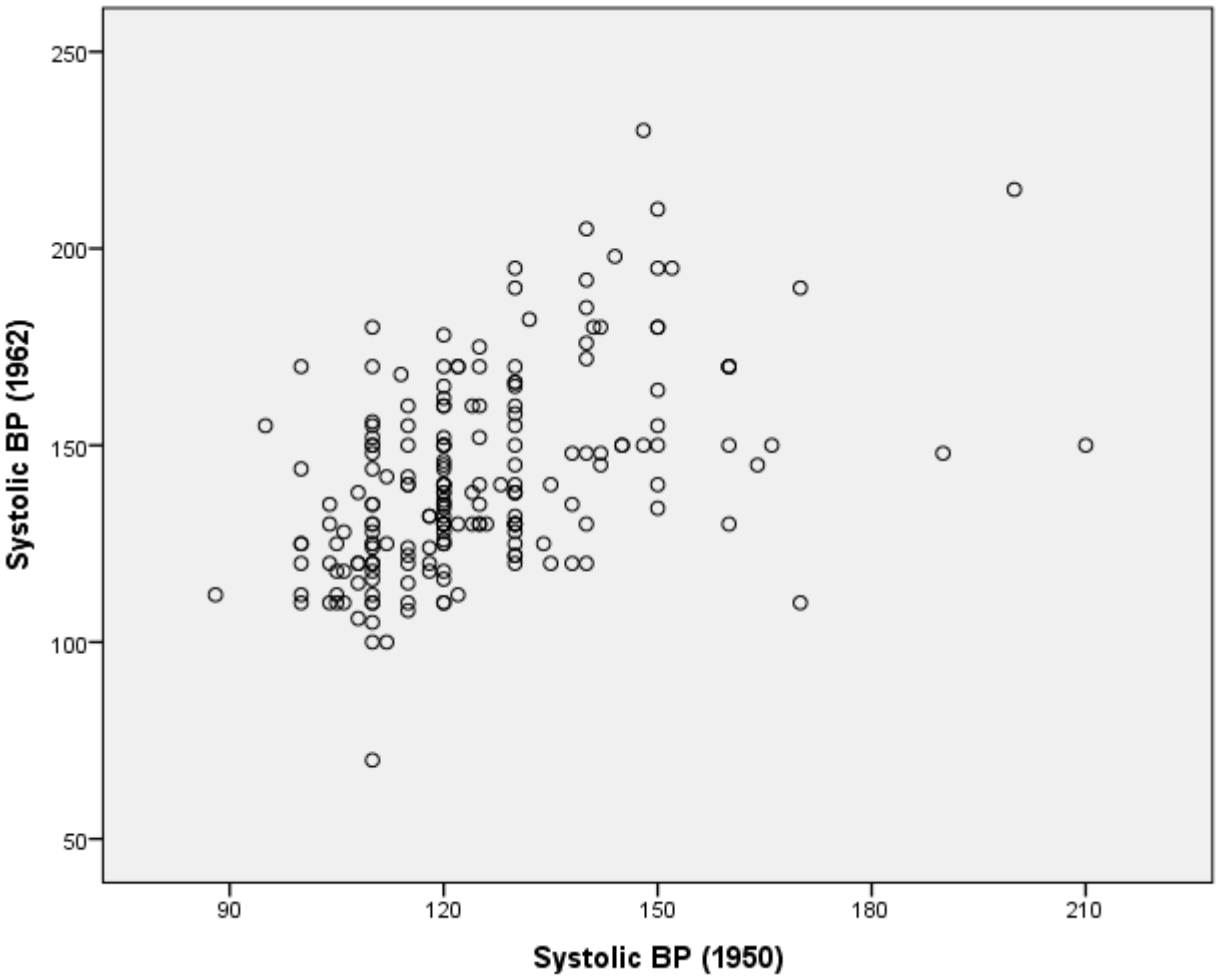
37252 Sample Exam

Question 1. (20 marks)

The analysis in this question is based on health data collected in 1950 from a sample of US adults, with follow-up data collected in 1962. There are two variables; the dependent variable is the systolic blood pressure (SBP) in 1962 of the respondents, while the independent variable is their SBP in 1950. As an initial analysis a scatterplot has been produced to look at the relationship between the two variables.



a) Based on the scatterplot, comment on the type, direction and strength of any relationship between the two variables.

(3 marks)

Notation used in Question 1:

- x is SBP in 1950
- y is SBP in 1962
- $\hat{y}$ is regression model prediction of SBP in 1962
- $\beta_0, \widehat{\beta_0}$ are the true and estimated values of model intercept parameters
- $\beta_1, \widehat{\beta_1}$ are the true and estimated values of model slope parameters

The plot shows SBP in 1950 and SBP in 1962 to be positively associated with a linear relationship. The clustering suggests a fairly strong relationship, but with increasing variance and a hint of negative curvature.

Following the scatterplot, a simple linear regression has been undertaken, which has produced the following output:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.486[a]	.236	.232	21.608

a. Predictors: (Constant), Systolic BP (1950)

b. Dependent Variable: Systolic BP (1962)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	28538.213	1	28538.213	61.123	.000[b]
	Residual	92445.707	198	466.898		
	Total	120983.920	199			

a. Dependent Variable: Systolic BP (1962)

b. Predictors: (Constant), Systolic BP (1950)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	90.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	59.043	10.654		5.542	.000	41.435	76.650
	Systolic BP (1950)	.661	.085	.486	7.818	.000	.521	.801

a. Dependent Variable: Systolic BP (1962)

b) Using the ANOVA Table, test whether the model is significantly better than a model with the intercept only. State clearly the null and alternative hypotheses of the test; and your conclusion based on the appropriate p-value.

(3 marks)

The null and alternative hypotheses are

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0.$$

The F-test statistic is 61.123, equating to a p-value reported as 0.000.

The null hypothesis can be rejected as $p < 0.05$.

As we have found statistical evidence of a non-zero slope, the regression model including the independent variable SBP in 1950 must be superior to that with constant only.

c) Write-down the fitted regression model for the relationship between SBP in1962 and SBP in1950. Interpret the value of the estimate for β_1 (the slope) for the regression line.

(2 marks)

The fitted regression model is given by

$$\hat{y} = 59.043 + 0.661x.$$

We can interpret $\widehat{\beta_0} = 59.043$ as the value of $\hat{y}$ when $x = 0$ (outside of the range of the sample data obviously).

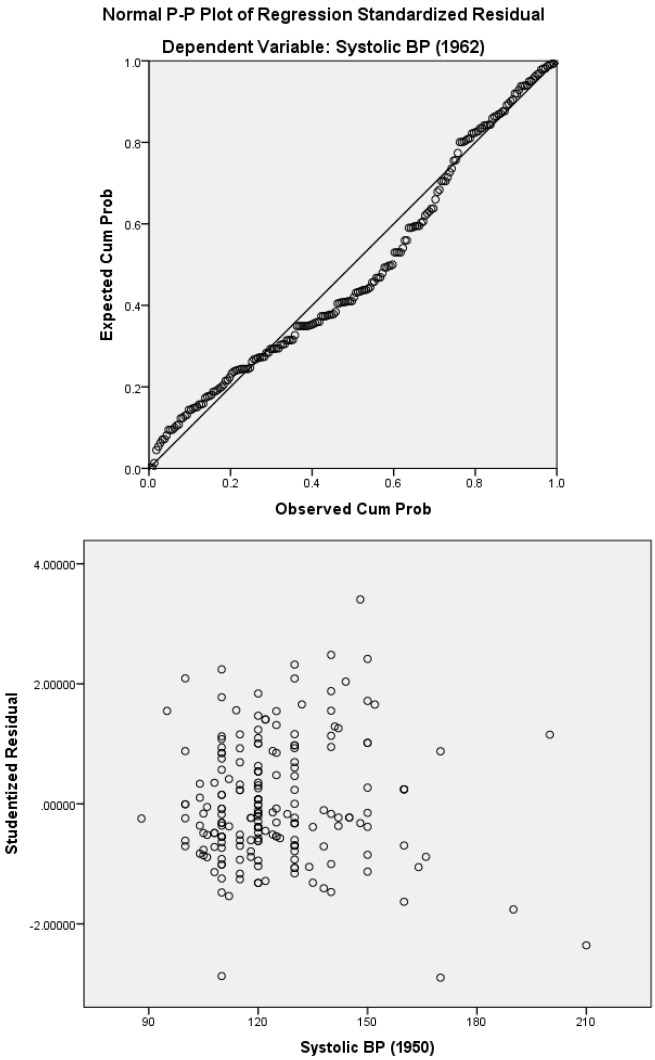
We can interpret $\widehat{\beta_1} = 0.661$ as the increase in $\hat{y}$ for a unit increase in x.

d) Find the 90% confidence interval for the intercept parameter β_0 and state clearly what a confidence interval means.

(3 marks)

The 90% CI for β_0 is $[41.435,76.650]$. So with 90% certainty we can say that $41.435 \leq \beta_0 \leq 76.650$, assuming the assumptions on the residuals are justifiable. Put another way, if we repeated the modelling using 100 different samples, we would **expect** 90% of the CIs to contain the true value β_0.

As part of the analysis, SPSS has also produced a normal P-P plot of the residuals and a scatterplot of the (internally) studentized residuals. These plots can be seen on the next page.



e) Discuss whether the two plots of the residuals support the assumptions of the regression model being satisfied.

(3 marks)

The P-P plot shows some departure from normality (as assumed for the residuals) between the 0.35-0.75 accumulated probability range. There is also obvious signs of increasing variance in the residuals and perhaps even some negative curvature, both contrary to independence and constant variance assumptions.

f) A colleague suggests transforming the data and fitting a model of the form

$$\frac{y_i}{\sqrt{x_i}} = \frac{\beta_0}{\sqrt{x_i}} + \beta_1\sqrt{x_i} + \frac{\varepsilon_i}{\sqrt{x_i}}.$$

Explain why this may help improve the fit of the model with respect to the assumptions of the regression model.

(2 marks)

By the definition of variance we have $var\left(\frac{\varepsilon_i}{\sqrt{x_i}}\right) = \frac{1}{x_i} var(\varepsilon_i)$, so if $var(\varepsilon_i) \propto x_i$, then $var\left(\frac{\varepsilon_i}{\sqrt{x_i}}\right) \propto 1$. In this case, the modified residuals will have constant variance (as assumed by the model).

The model suggested in part f) was fitted to the data by transforming the variables. This resulted in the following output for the transformed model.

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	90.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	one_sqrt_x	54.312	11.049	.383	4.915	.000	36.052	72.572
	SQRT(SBP_50)	.699	.089	.609	7.817	.000	.551	.847

a. Dependent Variable: Y over SQRT(X)

b. Linear Regression through the Origin

g) Interpret the fitted model parameters with respect to the original variables of SBP in 1962 and SBP in 1950. Comment on whether the model fits better in terms of how well the parameters are estimated. (*Hint: compare the standard errors for the model parameters from the two fitted models*.)

(4 marks)

(End of Question 1)

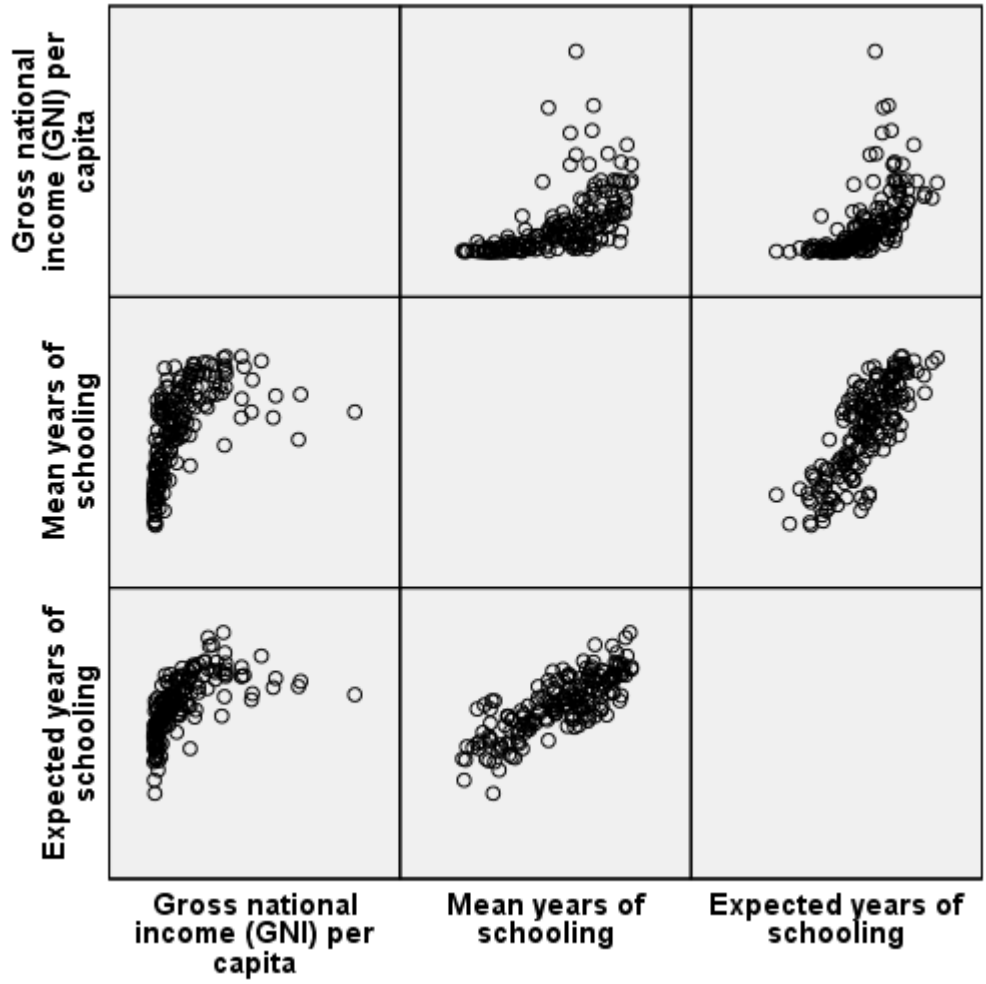
We can interpret $\widehat{\beta_0} = 54.312$ as the value of $\hat{y}$ when $x = 0$.
We can interpret $\widehat{\beta_1} = 0.699$ as the increase in $\hat{y}$ for a unit increase in x.

The parameter estimates are different but, interestingly, the standard errors of the estimates increased. So although this model has addressed, at least in part, the violation of the assumption of constant variance, this has not improved the fit of the model.

Question 2. (20 marks)

The analysis in this question uses multiple linear regression to explore the relationship between a country's Gross National Income per capita (GNI) and two measures of the education system; mean years of schooling and expected years of schooling.

Below is a matrix scatterplot looking at relationships between the three variables.



a) Discuss the strength and direction of the relationship between the dependent variable GNI and the two potential explanatory variables. State why using the log of GNI would be sensible when applying multiple linear regression.

(3 marks)

Notation used in Question 1:

- $school_M$ is mean years of schooling
- $school_E$ is expected years of schooling
- GNI is GNI per capita
- $\widehat{logGNI}$ is regression model prediction of the $\log_{10}$(GNI)
- $\beta_0, \widehat{\beta_0}$ are the true and estimated values of model intercept parameters
- $\beta_M, \widehat{\beta_M}$ are the true and estimated values of the coefficient associated with $school_M$

- $\beta_E, \widehat{\beta_E}$ are the true and estimated values of the coefficient associated with $school_E$
- $\beta_{M,ME}, \widehat{\beta_{M,ME}}$ are the true and estimated values of the coefficient associated with the interaction term $Mid\_East \times school_E$, with Mid_East a dummy variable for country group Middle East.

GNI is positively correlated with both $school_M$ and $school_E$. The relationship is non-linear which makes the strength of the relationships difficult to assess due to the scaling of the plots. The exponential-nature of the plots suggests modelling $logGNI$ as the dependent variable.

A multiple regression model is fitted with $\log_{10}$(GNI) as the dependent variable and including both explanatory variables. The output on the following page is created.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.825[a]	.681	.677	.29843

a. Predictors: (Constant), Expected years of schooling , Mean years of schooling

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	34.906	2	17.453	195.967	.000[b]
	Residual	16.387	184	.089		
	Total	51.294	186			

a. Dependent Variable: Log (base 10) of GNI per capita

b. Predictors: (Constant), Expected years of schooling , Mean years of schooling

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2.334	.108		21.689	.000		
	Mean years of schooling	.074	.012	.430	6.188	.000	.360	2.775
	Expected years of schooling	.081	.013	.440	6.339	.000	.360	2.775

a. Dependent Variable: Log (base 10) of GNI per capita

b) From the output, find and write-down the value of R-square and explain what this tells us in terms of how well the model fits.

(3 marks)

The coefficient of determination $R^2 = 0.681$, which is the proportion of Total Sum of Squares represented by Sum Square Regression . So 68.1% of the variation of

$\widehat{logGNI}$, defined as sum of squared deviations about its sample average, has been captured by the model, giving a quantification of the fit of the model to the sample data.

c) Write-down the fitted regression model.

(1 mark)

$$\widehat{logGNI} = 2.334 + 0.074\ school_M + 0.081\ school_E$$

d) For the fitted model, interpret the impact of both mean years of schooling and expected years of schooling on a country's GNI (*original scale*). Comment on the statistical significance of the parameters in relation to your interpretation.

(3 marks)

Holding $school_E$ constant, a unit increase in $school_M$ is associated with a $\widehat{\beta_M} = 0.074$ increase in $\widehat{logGNI}$ or $\widehat{GNI}$ to multiply by a factor of $10^{0.074} = 1.186$.

Holding $school_M$ constant, a unit increase in $school_E$ is associated with a $\widehat{\beta_E} = 0.081$ increase in $\widehat{logGNI}$ or $\widehat{GNI}$ to multiply by a factor of $10^{0.081} = 1.205$.

e) The output contains collinearity statistics. Explain why multi-collinearity is a problem for the interpretation of multiple regression models and comment on whether it is an issue in this particular model. You should refer back to an appropriate part of the matrix scatterplot.

(4 marks)

Multicollinearity is a problem caused by very high correlation between dependent variables included in the regression model. The effect of this is to increase the errors of the coefficient estimates. It also causes difficulty in analysis of the effects of changes in the independent variables as they move together.

The degree of multicollinearity can be quantified by Variance Inflation Factors (VIF). The threshold value we look for is $VIF_i > 10$, where VIF_i refers to the i-th sample point.

To extend the analysis, a categorical variable grouping countries into broad regions is included. The regions are; Europe, Americas, Oceania, Middle East, Asia, Africa. Choosing Europe as the reference group, the following SPSS output for the estimated model parameters is created.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.517	.146		17.212	.000
	Mean years of schooling	.069	.013	.405	5.540	.000
	Expected years of schooling	.072	.012	.391	6.107	.000
	Americas	.011	.067	.008	.171	.864
	Oceanea	-.352	.097	-.151	-3.607	.000
	Mid_East	.289	.087	.150	3.335	.001
	Asia	.005	.072	.003	.064	.949
	Africa	-.145	.082	-.124	-1.774	.078

a. Dependent Variable: Log (base 10) of GNI per capita

f) Interpret the impact of each category of country group (*relative to Europe*) on a country's GNI. Comment on the statistical significance of the parameters for each category in relation to your interpretation.

(4 marks)

$\widehat{GNI}$ for the Americas is $10^{0.011} = 1.026$ times that of Europe.
etc…

The continuous independent variables are both highly significant.

Of the five dummy variables for country group, only those for Oceania and Mid_East are significant, with those for the Americas and Asia highly non-significant.

A further extension has included the interaction between mean years of schooling and countries in the Middle East.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.487	.149		16.639	.000
	Mean years of schooling	.073	.013	.423	5.614	.000
	Expected years of schooling	.072	.012	.390	6.087	.000
	Americas	.019	.067	.014	.277	.782
	Oceanea	-.346	.098	-.149	-3.548	.000
	Mid_East	.566	.294	.293	1.925	.056
	Asia	.015	.073	.010	.200	.841
	Africa	-.127	.083	-.109	-1.521	.130
	Mean Schooling by Middle East	-.034	.034	-.144	-.985	.326

a. Dependent Variable: Log (base 10) of GNI per capita

g) Test whether the interaction effect is significant stating clearly your hypotheses, test statistic, p-value, and conclusion.

(2 marks)

The null and alternative hypotheses are

$$H_0: \beta_{M,ME} = 0$$
$$H_A: \beta_{M,ME} \neq 0.$$

The t-test statistic is -0.985, equating to a p-value of 0.326. At our preferred significance level $\alpha = 0.05$ the null hypothesis cannot be rejected. So there is no statistical evidence that $\widehat{logGNI}$ has different sensitivities to changes in $school_M$ for countries in the Middle East compared to countries elsewhere.

Question 3.

Sex * Low Pay Indicator Crosstabulation

			Low Pay Indicator		
			.00	1.00	Total
Sex	male	Count	1342	163	1505
		Expected Count	1255.4	249.6	1505.0
		% within Sex	89.2%	10.8%	100.0%
	female	Count	1268	356	1624
		Expected Count	1354.6	269.4	1624.0
		% within Sex	78.1%	21.9%	100.0%
Total		Count	2610	519	3129
		Expected Count	2610.0	519.0	3129.0
		% within Sex	83.4%	16.6%	100.0%

a) Using the percentages in the cross-tabulation, describe the relationship between gender and receiving low pay.

Using the percentages, males are less likely to receive low pay (10.8%) compared to females (21.9%).

b) Using the percentages (or counts) calculate the odds of males receiving low pay and the odds for females. Hence calculate (and interpret) the odds ratio of receiving low pay for females relative to males.

Odds (Sex=male) = 163 / 1342 = 0.12146
Odds (Sex=female) = 356 / 1268 = 0.280757
Odds Ratio = 0.280757 / 0.12146 = 2.312

The odds for females receiving low pay are 2.31 times higher than the odds for males receiving low pay.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	69.444 [a]	1	.000		

c) Using the Chi-Square test of association, decide whether there is a significant association between gender and receiving low pay. Make sure you:
- state clearly your null and alternative hypothesis,
- explain the role of the expected counts in the calculation of the test statistic,
- state the value of your test statistic with its associated p-value, and your conclusion with respect to the hypotheses.

Let LPI be Low Pay Indicator
H_0: Sex and LPI are independent
H_A: Sex and LPI are not independent

The expected counts are what the cells would look like IF the margins were distributed based on no association and therefore we can compare the observed to these expected counts under the null hypothesis.

The test statistic is 69.444 (p-value less than 0.0005) so we reject the null hypothesis of independence and conclude an association exists between the variables .

d) For the fitted model, show that the estimated odds of low pay from the model are given by

$$\widehat{odds} = e^{(\hat{\beta}_0 + \hat{\beta}_1 \times Sex)}$$

and the fitted probabilities are given by

$$\hat{p} = \frac{e^{(\hat{\beta}_0 + \hat{\beta}_1 \times Sex)}}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1 \times Sex)}}$$

where the dummy variable $Sex = 0$ for males, $Sex = 1$ for females and $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of β_0 and β_1.

The logistic regression model for predicted log-odds $\widehat{\log\_odds}$, is given by
$\widehat{\log\_odds} = log\frac{\hat{p}}{1-\hat{p}} = \hat{\beta}_0 + \hat{\beta}_1 \times Sex.$

Taking the exponential of both sides gives the model in odds-space

$$\frac{\hat{p}}{1-\hat{p}} = e^{\hat{\beta}_0 + \hat{\beta}_1 \times Sex}$$

which after solving for $\hat{p}$ is

$$\hat{p} = \frac{e^{(\hat{\beta}_0 + \hat{\beta}_1 \times Sex)}}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1 \times Sex)}}.$$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 [a]	sex	.838	.102	67.006	1	.000	2.312
	Constant	-2.108	.083	645.971	1	.000	.121

a. Variable(s) entered on step 1: sex.

e) From the output, write down the fitted model in terms of the Ln(odds) of receiving low pay. Does the output support a significant relationship between gender and receiving low pay? Make sure you justify your answer with a hypothesis test and associated p-value.

The model for log-odds is

$$\widehat{\log\_odds} = log\frac{\hat{p}(Sex)}{1-\hat{p}(Sex)} = -2.108 + 0.838 \times Sex$$

or

$$\widehat{\log\_odds} = log\frac{\hat{p}(0)}{1-\hat{p}(0)} = -2.108$$

for males and

$$\widehat{\log\_odds} = log\frac{\hat{p}(1)}{1-\hat{p}(1)} = -1.27$$

for females.

To test the significance of Sex define the hypotheses

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0.$$

The Wald test statistic is 67.006 (p-value less than 0.0005) so the null hypothesis is rejected and the conclusion drawn of different log-odds for males and females.

f) Using the relationships given in part d) (or otherwise), show that the binary logistic model gives fitted values for the odds for males and females receiving low pay that match those calculated in part b), and hence that the odds ratio for females relative to males is 2.312.

Odds for males is $\frac{\hat{p}(0)}{1-\hat{p}(0)} = e^{-2.108} = 0.12148.$

Odds for females is $\frac{\hat{p}(1)}{1-\hat{p}(1)} = e^{-2.108+0.838} = 0.28083.$

So the odds ratio (female to male) is $\frac{0.28083}{0.12148} = 2.312.$

Question 4.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	325.110	5	.000
	Block	325.110	5	.000
	Model	325.110	5	.000

a) Using the appropriate information, discuss whether there is evidence that the overall model fits better than a model with just the intercept. State the value of the test statistic and the associated p-value that supports your answer.

The Omnibus Tests of Model Coefficients show us that the overall model fits significantly better than just the intercept. This is supported by a test statistic of 325.110 and a p-value less than 0.05.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 [a]	age	-.318	.024	174.409	1	.000	.728
	age_sq	.004	.000	142.813	1	.000	1.004
	sex	.876	.108	66.170	1	.000	2.402
	col_reg			33.214	2	.000	
	col_reg(1)	.654	.127	26.392	1	.000	1.923
	col_reg(2)	.810	.159	25.967	1	.000	2.249
	Constant	3.543	.433	66.822	1	.000	34.556

a. Variable(s) entered on step 1: age, age_sq, sex, col_reg.

b) Using odds ratios interpret the **linear relationship** between age and low pay. State whether the relationship is statistically significant based on appropriate p-value.

A one unit increase in age is associated with the odds of low pay multiplies by a factor of 0.728. The effect is highly significant as the Wald statistic 174.409 has a p-value less than 0.05.

c) Using the odds ratio, interpret the relationship between gender and low pay.

Relative to males, being female multiplies the odds of low pay by a factor 2.402.

Categorical Variables Codings

		Frequency	Parameter coding (1)	Parameter coding (2)
Collapsed Region	South East	944	.000	.000
	Rest of England	1665	1.000	.000
	Rest of UK	520	.000	1.000

d) Using odds ratios, interpret the relationship between 'region of residence' and low pay. State whether the two odds ratios are statistically significant based on appropriate p-values.

Relative to the 'South East' category, living in rest of England multiplies the odds of low pay by a factor 1.923.
Relative to the 'South East' category, living in rest of UK multiplies the odds of low pay by a factor 2.249.
In both cases the individual odds ratios are highly significant but so is the overall test for relationship (all p-values less than 0.05).

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	21.400	8	.006

e) Explain why the output for the Hosmer and Lemeshow Test supports exploring interactions between the variables in the model.

This test has the null hypothesis that the patterns in the fitted probabilities match well to the patterns in the observed probabilities, while the alternative says the patterns are significantly different. Therefore, for a given set of X's we can see if we have the right structure to our model for that set of X's. In this case we reject the null so our current model structure is not explaining the observed patterns very well, hence we should try interactions.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 [a]	age	-.318	.024	174.279	1	.000	.728
	age_sq	.004	.000	142.620	1	.000	1.004
	sex	.746	.225	11.026	1	.001	2.109
	col_reg			9.155	2	.010	
	col_reg(1)	.524	.216	5.905	1	.015	1.689
	col_reg(2)	.750	.259	8.360	1	.004	2.116
	col_reg * sex			.581	2	.748	
	col_reg(1) by sex	.196	.266	.543	1	.461	1.217
	col_reg(2) by sex	.088	.327	.072	1	.788	1.092
	Constant	3.629	.453	64.089	1	.000	37.691

a. Variable(s) entered on step 1: age, age_sq, sex, col_reg, col_reg * sex.

f) Is the interaction term significant? State the value of the test statistic and the associated p-value that supports your answer.

No it is NOT significant. The overall test has a value 0.581 with a p-value of 0.748 with and each separate interaction term begin insignificant as well.

g) For an individual aged 40 years, work-out the six fitted probabilities associated with combinations of gender and region of residence. (**Hint:** *Write-out the model for the six combinations and then use the formula in part d) of Question 4.*)

Allow either answer with the interaction

Male by SE: $\text{Ln(odds)} = 3.629 - 0.318\times40 + 0.004\times40^2 \Rightarrow p = 0.064$

Male by rest Eng: $\text{Ln(odds)} = 3.629 - 0.318\times40 + 0.004\times40^2 + 0.524 \Rightarrow p = 0.103$

Male by rest UK: $\text{Ln(odds)} = 3.629 - 0.318\times40 + 0.004\times40^2 + 0.750 \Rightarrow p = 0.126$

Female by SE: $\text{Ln(odds)} = 3.629 - 0.318\times40 + 0.004\times40^2 + 0.746 \Rightarrow p = 0.125$

Female by rest Eng: $\text{Ln(odds)} = 3.629 - 0.318\times40 + 0.004\times40^2 + 0.746 + 0.524 + 0.196 \Rightarrow p = 0.227$

Female by rest UK: $\text{Ln(odds)} = 3.629 - 0.318\times40 + 0.004\times40^2 + 0.746 + 0.750 + 0.088 \Rightarrow p = 0.248$

OR without

Male by SE: $\text{Ln(odds)} = 3.543 - 0.318\times40 + 0.004\times40^2 \Rightarrow p = 0.059$ (1)

Male by rest Eng: $\text{Ln(odds)} = 3.543 - 0.318\times40 + 0.004\times40^2 + 0.654 \Rightarrow p = 0.107$

Male by rest UK: $\text{Ln(odds)} = 3.543 - 0.318\times40 + 0.004\times40^2 + 0.810 \Rightarrow p = 0.123$

Female by SE: $\text{Ln(odds)} = 3.543 - 0.318\times40 + 0.004\times40^2 + 0.876 \Rightarrow p = 0.130$

Female by rest Eng: $\text{Ln(odds)} = 3.543 - 0.318\times40 + 0.004\times40^2 + 0.876 + 0.654 \Rightarrow p = 0.223$

Female by rest UK: $\text{Ln(odds)} = 3.543 - 0.318\times40 + 0.004\times40^2 + 0.876 + 0.810 \Rightarrow p = 0.251$