MATH 204 - SOLUTIONS 4

1. Given that

$$\boldsymbol{X} = \left[\begin{array}{cccc} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{array} \right]^\top$$

we have that, by the multiplication rules given

$$\boldsymbol{X}^\top \boldsymbol{X} = \left[\begin{array}{cc} n & S_x \\ S_x & S_{xx} \end{array} \right]$$

where

$$S_x = \sum_{i=1}^n x_i \qquad S_{xx} = \sum_{i=1}^n x_i^2$$

The matrix inverse is computed by using the result given on the handout; a square $k \times k$ matrix A has an **inverse**, denoted A^{-1} if

$$A.A^{-1} = A^{-1}.A = I_k$$

Here we set $A = \boldsymbol{X}^\top \boldsymbol{X}$. We need to find the four constants $a_{11}, a_{12}, a_{21}, a_{22}$ such that

$$\left[\begin{array}{cc} n & S_x \\ S_x & S_{xx} \end{array} \right] \left[\begin{array}{cc} a_{11} & a_{12} \\ a_{21} & a_{22} \end{array} \right] = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right]$$

Thus we have the four simultaneous equations to solve

$$\begin{array}{lrcrcl} (1) & na_{11} & + & S_x a_{21} & = & 1 \\ (2) & na_{12} & + & S_x a_{22} & = & 0 \\ (3) & S_x a_{11} & + & S_{xx} a_{21} & = & 0 \\ (4) & S_x a_{12} & + & S_{xx} a_{22} & = & 1 \end{array}$$

After some manipulation, we find that

$$a_{11} = \frac{S_{xx}}{nS_{xx} - S_x S_x} \qquad a_{12} = a_{21} = \frac{-S_x}{nS_{xx} - S_x S_x} \qquad a_{22} = \frac{n}{nS_{xx} - S_x S_x}$$

so that

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \frac{1}{nS_{xx} - S_x S_x} \left[\begin{array}{cc} S_{xx} & -S_x \\ -S_x & n \end{array} \right]$$

Note that in general for 2×2 matrices, we have the general formula

$$\left[\begin{array}{cc} a & b \\ c & d \end{array} \right]^{-1} = \frac{1}{ad - bc} \left[\begin{array}{cc} d & -b \\ -c & a \end{array} \right]$$

provided that $ad - bc \neq 0$. Finally, we have that

$$\boldsymbol{X}^\top \underset{\sim}{y} = \left[\begin{array}{c} S_y \\ S_{xy} \end{array} \right]$$

where

$$S_y = \sum_{i=1}^n y_i \qquad S_{xy} = \sum_{i=1}^n x_i y_i$$

and hence, by multiplying out, we get

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \underset{\sim}{y} = \left[\begin{array}{c} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{array} \right]$$

where

$$\widehat{\beta}_0 = \frac{S_{xx}S_y - S_x S_{xy}}{nS_{xx} - S_x S_x} \qquad \widehat{\beta}_1 = \frac{nS_{xy} - S_x S_y}{nS_{xx} - S_x S_x}$$

Now note that

$$\frac{nS_{xy} - S_x S_y}{nS_{xx} - S_x S_x} = \frac{S_{xy} - \dfrac{S_x S_y}{n}}{S_{xx} - \dfrac{S_x S_x}{n}} = \frac{SS_{xy}}{SS_{xx}}$$

where

$$SS_{xy} = S_{xy} - \frac{S_x S_y}{n} = S_{xy} - n\,\overline{x}\,\overline{y} = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

and similarly

$$SS_{xx} = S_{xx} - \frac{S_x S_x}{n} = \sum_{i=1}^{n}(x_i - \overline{x})^2.$$

These results use the shortcut formula for sample variance given on page 69 of McClave and Sincich. Thus the formula for $\widehat{\beta}_1$ matches the one given in lectures. A similar calculation verifies the result for $\widehat{\beta}_0$.

2. For this problem, we use ANOVA and linear regression techniques, specifically multiple regression. Note that **Model** and **Vendor** are factor predictors, so we use the `General Linear Model` pulldown menu in SPSS.

   The SPSS output for a series of models is attached; we fit in turn each of the single predictor models, then the multiple regression model with all variables included, then different models with variables and interactions included. We use inspection of p-values in ANOVA tables and R^2 statistics to assess the most suitable model fit. For the analysis, price is in thousands of pounds.

   Note that this is only an informal model comparison procedure; we do not use the formal ANOVA-F test comparison models developed later.

   Our conclusions are summarized as follows:

   - In the main effects only models (Models 1 - 4), **Model**, **Age**, and **Mileage** are important predictors, as all have significant p-values in the one-way ANOVA. Of these variables, **Model** seems to be the most important predictor, with an R^2 value of 0.77. The variable **Vendor** is not significant at the $\alpha = 0.05$ significance level ($p = 0.089$).
   - In the multiple regression model with interaction between the two factor predictors (Model 5), **Age** and **Model** appear to be significant predictors (precise interpretation may be difficult in this unbalanced design). The R^2 value is now 0.947, indicating good explanatory power.
   - After checking a selection of models (Model 6 - 10) it seems that the best model in terms of simplicity and good explanatory power is the model

     Age $+$ **Model**

     No other terms appear to be significant, and also $R^2 = 0.906$ with Adjusted $R^2 = 0.896$, so the explanatory power is good.
   - Inspection of the residuals indicates that overall the model assumptions are met, as we see no pattern in the residuals. There may be evidence of a single outlier (the car with the highest observed price)
   - Inspection of the parameter estimates indicates that price **decreases** with increasing **Age** (estimated coefficient is -1.079, standard error 0.138), and that the 500 series (**Model**=0) has the highest price, with coefficient 13.486+11.966 = 25.452.

SPSS Output for Exercises 4 Q2

Model 1: Mod

Dependent Variable: Price (1000 GBP)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1105.468(a)	4	276.367	45.279	.000
Intercept	11607.038	1	11607.038	1901.661	.000
Mod	1105.468	4	276.367	45.279	.000
Error	299.078	49	6.104		
Total	13658.417	54			
Corrected Total	1404.546	53			

a R Squared = .787 (Adjusted R Squared = .770)

Dependent Variable: Price (1000 GBP)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	9.236	.618	14.953	.000	7.994	10.477
[Mod=0]	12.843	1.070	12.005	.000	10.693	14.993
[Mod=1]	5.610	1.266	4.432	.000	3.067	8.154
[Mod=2]	9.922	.996	9.963	.000	7.921	11.923
[Mod=3]	5.648	.888	6.361	.000	3.863	7.432
[Mod=4]	0(a)	.	.	.	.	.

a This parameter is set to zero because it is redundant.

Model 2: Age

Dependent Variable: Price (1000 GBP)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	258.133(a)	1	258.133	11.709	.001
Intercept	4109.494	1	4109.494	186.402	.000
Age	258.133	1	258.133	11.709	.001
Error	1146.413	52	22.046		
Total	13658.417	54			
Corrected Total	1404.546	53			

a R Squared = .184 (Adjusted R Squared = .168)

Dependent Variable: Price (1000 GBP)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	19.409	1.422	13.653	.000	16.557	22.262
Age	-1.128	.330	-3.422	.001	-1.790	-.467

Model 3: Mile

Dependent Variable: Price (1000 GBP)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	326.165(a)	1	326.165	15.728	.000
Intercept	5063.081	1	5063.081	244.144	.000
Mile	326.165	1	326.165	15.728	.000
Error	1078.381	52	20.738		
Total	13658.417	54			
Corrected Total	1404.546	53			

a R Squared = .232 (Adjusted R Squared = .217)

Dependent Variable: Price (1000 GBP)

					95% Confidence Interval	
Parameter	B	Std. Error	t	Sig.	Lower Bound	Upper Bound
Intercept	19.302	1.235	15.625	.000	16.823	21.781
Mile	-.209	.053	-3.966	.000	-.315	-.103

Model 4: Vend

Dependent Variable: Price (1000 GBP)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	209.561(a)	4	52.390	2.148	.089
Intercept	12329.637	1	12329.637	505.573	.000
Vend	209.561	4	52.390	2.148	.089
Error	1194.985	49	24.387		
Total	13658.417	54			
Corrected Total	1404.546	53			

a R Squared = .149 (Adjusted R Squared = .080)

Dependent Variable: Price (1000 GBP)

					95% Confidence Interval	
Parameter	B	Std. Error	t	Sig.	Lower Bound	Upper Bound
Intercept	13.503	1.370	9.859	.000	10.751	16.256
[Vend=0]	3.015	2.023	1.490	.143	-1.050	7.081
[Vend=1]	5.054	2.219	2.278	.027	.595	9.514
[Vend=2]	1.925	2.141	.899	.373	-2.378	6.229
[Vend=3]	-.511	1.937	-.264	.793	-4.403	3.382
[Vend=4]	0(a)	.	.	.	.	.

a This parameter is set to zero because it is redundant.

Model 5: Age + Mile + Mod + Vend + Mod.Vend

Dependent Variable: Price (1000 GBP)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1329.511(a)	24	55.396	21.410	.000
Intercept	1907.237	1	1907.237	737.122	.000
Age	47.504	1	47.504	18.360	.000
Mile	1.769	1	1.769	.684	.415
Mod	604.015	4	151.004	58.361	.000
Vend	14.839	4	3.710	1.434	.248
Mod * Vend	36.082	14	2.577	.996	.482
Error	75.035	29	2.587		
Total	13658.417	54			
Corrected Total	1404.546	53			

a R Squared = .947 (Adjusted R Squared = .902)

Model 6: Age + Mile + Mod + Vend

Dependent Variable: Price (1000 GBP)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1293.428(a)	10	129.343	50.053	.000
Intercept	2413.866	1	2413.866	934.113	.000
Mod	888.417	4	222.104	85.949	.000
Vend	16.608	4	4.152	1.607	.190
Age	60.368	1	60.368	23.361	.000
Mile	2.461	1	2.461	.952	.335
Error	111.117	43	2.584		
Total	13658.417	54			
Corrected Total	1404.546	53			

a R Squared = .921 (Adjusted R Squared = .902)

Model 7: Age + Mod + Vend

Dependent Variable: Price (1000 GBP)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1290.967(a)	9	143.441	55.569	.000
Intercept	2474.277	1	2474.277	958.528	.000
Mod	927.675	4	231.919	89.845	.000
Vend	18.131	4	4.533	1.756	.155
Age	123.195	1	123.195	47.726	.000
Error	113.579	44	2.581		
Total	13658.417	54			
Corrected Total	1404.546	53			

a R Squared = .919 (Adjusted R Squared = .903)

Model 8: Age + Mod

Dependent Variable: Price (1000 GBP)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1272.836(a)	5	254.567	92.774	.000
Intercept	2949.842	1	2949.842	1075.032	.000
Mod	1014.703	4	253.676	92.449	.000
Age	167.368	1	167.368	60.995	.000
Error	131.710	48	2.744		
Total	13658.417	54			
Corrected Total	1404.546	53			

a R Squared = .906 (Adjusted R Squared = .896)

Model 9: Age + Mile + Mod

Dependent Variable: Price (1000 GBP)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1276.820(a)	6	212.803	78.307	.000
Intercept	2953.826	1	2953.826	1086.941	.000
Mod	920.691	4	230.173	84.698	.000
Age	61.768	1	61.768	22.729	.000
Mile	3.985	1	3.985	1.466	.232
Error	127.725	47	2.718		
Total	13658.417	54			
Corrected Total	1404.546	53			

a R Squared = .909 (Adjusted R Squared = .897)

Model 10: Age + Mod + Mod . Age

Dependent Variable: Price (1000 GBP)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1292.291(a)	9	143.588	56.282	.000
Intercept	2147.345	1	2147.345	841.688	.000
Mod	270.552	4	67.638	26.512	.000
Age	160.470	1	160.470	62.899	.000
Mod * Age	19.455	4	4.864	1.906	.126
Error	112.254	44	2.551		
Total	13658.417	54			
Corrected Total	1404.546	53			

a R Squared = .920 (Adjusted R Squared = .904)

Final Model: Age + Mod

Dependent Variable: Price (1000 GBP)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1272.836(a)	5	254.567	92.774	.000
Intercept	2949.842	1	2949.842	1075.032	.000
Mod	1014.703	4	253.676	92.449	.000
Age	167.368	1	167.368	60.995	.000
Error	131.710	48	2.744		
Total	13658.417	54			
Corrected Total	1404.546	53			

a R Squared = .906 (Adjusted R Squared = .896)

Parameter Estimates

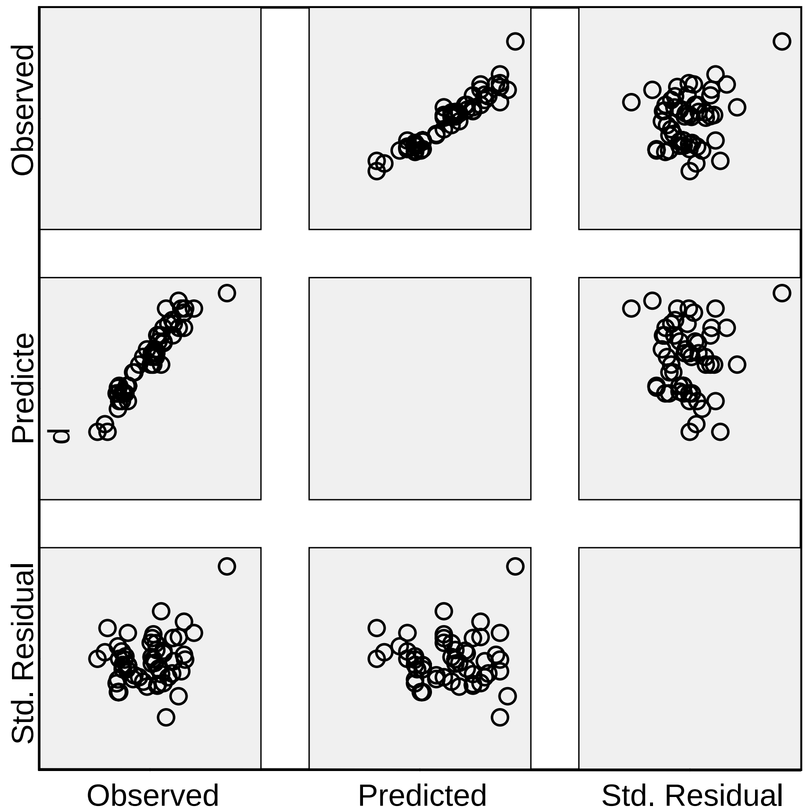
Dependent Variable: Price (1000 GBP)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	13.486	.684	19.720	.000	12.111	14.861
[Mod=0]	11.966	.726	16.482	.000	10.506	13.426
[Mod=1]	8.916	.948	9.401	.000	7.009	10.823
[Mod=2]	9.234	.674	13.709	.000	7.880	10.588
[Mod=3]	5.139	.599	8.582	.000	3.935	6.344
[Mod=4]	0(a)	.	.	.	.	.
Age	-1.079	.138	-7.810	.000	-1.357	-.802

a This parameter is set to zero because it is redundant.

Residuals

Dependent Variable: Price (1000 GBP)



Model: Intercept + Mod + Age