Name: David Levy

1. Data were collected from recently diagnosed breast cancer patients. The variables include the following:

	Variable	Description
c	age	Age at diagnosis, years
	educhi	Education (0 = less than college, 1 = college graduate)
	income	Income with 3 categories (low, moderate, high), coded using dummy variables inc2 (=1 for moderate and 0 otherwise) and inc3 (=1 for high and 0 otherwise)
	stage2or3	Cancer stage (0 = Stage 0 or 1, 1 = Stage 2 or 3)
	surgtype2cat	Type of surgery (0=lumpectomy, 1=mastectomy)
	ctq2cat	Childhood Trauma Questionnaire (0 = no childhood maltreatment, 1 = maltreatment)
	MDD	History of major depression (0 = no, 1 = yes)
c	monthsout	Time interval from diagnosis to survey, months
	married	Married (0=no, 1=yes)
c	MFSI	Fatigue score; higher values mean more fatigue

A new variable, age10, is created by dividing age by 10. A new variable monthsoutC is created by subtracting the mean value of monthsout (2 months) from monthsout. A multiple linear regression model is fit and provides the following output:

Table 1

Source	SS	df	MS
Model	1729.87781	7	247.125402
Residual	6966.43772	243	28.668468
Total	8696.31554	250	34.7852622

MFSI	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age10	-.7638744	.3036183	-2.52	0.013	-1.361934	-.1658148
educhi	-1.774245	.7371829	-2.41	0.017	-3.226329	-.3221614
stage2or3	-1.680167	.7222068	-2.33	0.021	-3.102751	-.2575825
surgtype2cat	1.331896	.8139114	1.64	0.103	-.2713257	2.935118
ctq2cat	2.340657	.7255173	3.23	0.001	.9115517	3.769763
MDD	1.740474	.8337702	2.09	0.038	.0981344	3.382813
monthsoutC	.7655608	.2913426	2.63	0.009	.1916816	1.33944
intercept ~~score~~	11.40696	1.894578	6.02	0.000	7.675068	15.13885

(a) (4 points) Calculate the value of R^2 for the model in Table 1 and provide an interpretation of its value. R^2, the coefficient of determination, quantifies the percent of variation in MFSI explained by its linear relationship with the regressors in the model. For this model, $R^2 = \frac{\text{Model SS}}{\text{Total SS}} = 1 - \frac{\text{RSS}}{\text{Total SS}} \approx 0.20$.

So about 20% of the variation in MFSI is explained by its linear relationship with the predictors selected for this model.

→ though it will be between 0 and 1, so we must multiply by 100 to get %

(b) (4 points) Calculate the value of the adjusted R^2 for the model in Table 1.

$$R^2_{adj} = 1 - \left[\frac{\left(\frac{RSS}{n-k-1}\right)}{\left(\frac{TotSS}{n-1}\right)}\right] = 1 - \left[\frac{(6966.43772/245)}{(8696.31554/250)}\right] = 1 - \left(\frac{28.668468}{34.7852622}\right) \approx 0.18$$

(c) (4 points) Calculate the value of the root mean squared error ($\hat{\sigma}$) for the model in Table 1.

$$MSE = \frac{\text{Residual SS}}{n-k-1}, \text{ where } k = \#\text{ predictors } (k+1 = \#\text{ parameters})$$

$$\sqrt{MSE} = \sqrt{\frac{\text{Residual SS}}{n-k-1}} = \sqrt{\frac{6966.43772}{(251)-(7+1)}} = 5.3543$$

We can also locate $\hat{\sigma}^2$ (MSE) in the table, and simply take the square root.

(d) (4 points) Calculate the F value for the overall F test and provide its distribution under the null hypothesis. What do you conclude if the null hypothesis is rejected?

$$F = \left(\frac{\text{Reg MS}}{MSE}\right) \sim F_{k,n-k-1} \quad \text{so} \quad F = \left(\frac{247.125402}{28.668468}\right) = 8.62011 \sim F_{7,243}$$

This is an omnibus F-test with a null hypothesis that all the regression coefficients except the intercept equal zero. Rejecting the null would mean at least one of these coefficients is not zero.

(e) (4 points) Provide a quantitative interpretation of the regression coefficient for age10.

For a ~~0.1~~ 10-year increase in age at diagnosis, there is, on average, a decrease in mean fatigue score of about 0.76 points, controlling for all of the other predictors in the model (i.e. holding them constant). -2

(f) (4 points) Provide a quantitative interpretation of the regression coefficient for monthsoutC.

Holding all other predictors constant, for a one-month increase from the average time interval from diagnosis to survey, there is, on average, an increase in mean fatigue score of about 0.77 points.

Your colleague wants to test for an interaction between history of major depression and marital status. She creates a new variable MDDxmarried as the product of MDD and married, and fits the following model:

Table 2

MFSI	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age10	-.6876825	.3018588	-2.28	0.024	-1.282288	-.0930765
educhi	-1.644876	.731026	-2.25	0.025	-3.084863	-.2048904
stage2or3	-1.604222	.7150399	-2.24	0.026	-3.012718	-.1957256
surgtype2cat	1.268076	.805516	1.57	0.117	-.3186412	2.854794
ctq2cat	2.515483	.7210301	3.49	0.001	1.095187	3.935779
MDD	-.4167892	1.189343	-0.35	0.726	-2.759576	1.925997
monthsoutC	.7040996	.2892268	2.43	0.016	.1343763	1.273823
MDDxmarried	3.623243	1.43924	2.52	0.012	.7882069	6.45828
_cons	10.88522	1.885531	5.77	0.000	7.171078	14.59937

(g) (5 points) Has your colleague fit an appropriate model for testing for an interaction between history of major depression and marital status? Explain your answer.

Certainly not. This model is not hierarchically well-structured because predictors married and MDD are considered lower-order terms when their interaction term (which is considered higher-order) is included in the model. MDD is included, but married is not, so this is not an appropriate model for testing this interaction.

(h) (5 points) You decide to model age using a cubic regression spline with one knot placed at age 60 years. What variables to you need to create and include in the regression model to model age using a cubic spline?

$x_{i1} \equiv x_i$

$x_{i2} \equiv \begin{cases} 0 & \text{for } x_i \leq 60 \\ x_i - c_i & \text{for } x_i > 60 \end{cases}$

X

You regress MFSI on the two dummy variables for income and obtain the following output:

Table 3

Source	SS	df	MS
Model	150.564005	2	75.2820024
Residual	8715.47885	263	33.1387029
Total	8866.04286	265	33.4567655

Number of obs	= 266
F(2, 263)	= 2.27
Prob > F	= 0.1052
R-squared	= 0.0170
Adj R-squared	= 0.0095
Root MSE	= 5.7566

$\hat{Y}$ MFSI	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
β_1 inc2	-1.010645	1.058237	-0.96	0.340	-3.09434	1.073051
β_2 inc3	-1.798058	.8494615	-2.12	0.035	-3.470669	-.1254467
β_0 _cons	8.859701	.7032834	12.60	0.000	7.474919	10.24448

$y = \beta_0 + \beta_1 inc2 + \beta_2 inc3$

(i) Explain how you would test the following hypotheses. If you are able to conduct the test, do so. If you do not have the information required to conduct the test, explain what additional information you need.

i. (4 points) The mean MFSI for women with moderate income is equal to the mean MFSI for women with low income.

$H_0: \beta_1 = 0 \; ; \; H_1: \beta_1 \neq 0$

Because then there would be no difference. This test is included in the table, as a t-statistic and p-value (can also be done w/ partial F-test) $p = 0.340$, so null cannot be rejected.

ii. (4 points) The mean MFSI for women with moderate income is equal to the mean MFSI for women with high income.

$H_0: \beta_2 = 0$; $H_1: \beta_2 \neq 0$, again included in the table, and again can be done with a partial F-test.

$p = 0.035$. We can reject the null hypothesis at the $\alpha = 0.05$ confidence level. −3

$H_0: \beta_2 = \beta_3$

iii. (4 points) The mean MFSI does not differ by income category.

$H_0: \beta_1 = \beta_2 = 0$ (Omnibus F-test!) Included in table already ✓

H_1: "At least one is not 0"

$$F = \frac{\text{Reg MS}}{\text{MSE}} = \frac{75.2820024}{33.1387029} = 2.272 \sim F_{2,263}$$

$p = 0.1052$, cannot reject null hypothesis

2. Suppose we have data that follow the model $Y_i = \beta_0 + \beta_0 x_i + \epsilon_i$ where $Var(\epsilon_i)$ is proportional to x_i^2 and the ϵ_i's are independent.

(a) (5 points) Explain the main advantage of using weighted least squares (WLS) to obtain estimates of the regression coefficients β_0 and β_1 in this data setting. −1

Essentially, a WLS approach entails weighting each observation by the inverse of its variance, so that observations with low variance (reliable information) are weighted more heavily, while observations with high variance (not-so-reliable information) are given less weight. The thought is that this makes our estimates of the regression coefficients better.

vague

(b) (4 points) To conduct a WLS regression, we express the sum of squared deviations as $\Sigma_{i=1}^{n} w_i[y_i - (\beta_0 + \beta_1 x_i)]^2$. What are the appropriate weights w_i?

Because $Var(\epsilon_i) \propto x_i^2$, $Var(Y) \propto x_i^2$, so we should use ✓ a weight of $\frac{1}{k x_i^2}$, where k is the appropriate proportionality constant.

3. Answer true or false and provide a brief explanation.

(a) (4 points) The regression to the mean effect is stronger when the variables involved are more highly correlated. ✓

False. Higher correlation will mitigate the RTM effects. We can quantify this by observing $\frac{\hat{y}_i - \bar{Y}}{s_y} = r_{xy} \frac{\hat{x}_i - \bar{X}}{s_x}$ and that r_{xy} is bounded by $0 \leq r_{xy} \leq 1$. Also percent RTM can be quantified by $(100)(1 - r_{xy})$, so high correlation yields low effect.

(b) (4 points) In order for a variable D to be considered a confounder of the relationship

between X and Y, D must be a risk factor for Y.

True. We defined confounding as a "lack of comparability between exposure groups" due to an inherent risk for the same outcome (Y) caused by another, unaccounted for factor (D in this example). ✓

(c) (4 points) A major use of linear regression is to adjust for confounding.

True. Much of the time, "controlling" for particular variables helps us to remove confounding effects. ✓

(d) (4 points) Two predictors that are highly correlated are likely to interact.

Not necessarily. There can be highly correlated predictors that do not interact and predictors that interact, but are not highly correlated. ✓

(e) (4 points) An unusual observation is more likely to be influential on the regression coefficients in a small data set than in a large data set. In general, this is true, but we need to be far more specific about what unusual means. X-outlier? ✓ conditional Y-outlier? Then definitely yes. Marginal y-outlier? Possibly, but we can't say for sure unless we see information about its residuals, leverage, etc.

(f) (4 points) When comparing two models, the model with the higher R^2 is expected to have better predictive performance on test data (that is, data not used to develop the models).

Not necessarily, no. R^2 can be sensitive even to just the number of predictors used (we will never reduce R^2 by adding more predictors). But based on our methods, we could train a model with high R^2, then get horrible predictions because we were overfitting. ✓

4. Suppose that we have data that follow the model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, $E(\epsilon) = \mathbf{0}$, $Var(\epsilon) = \sigma^2\mathbf{I}$.

(a) (4 points) In this model, what is $E(\mathbf{Y})$? Show your work.

$$E(Y) = E(X\beta + \epsilon) = E(X\beta) + \underbrace{E(\epsilon)}_{0} = X\beta$$ ✓

(b) (4 points) What is the expected value of the least squares estimator of the regression

coefficients, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$? Show your work.

$E(\hat{\beta}) = E[(X'X)^{-1}X'Y] = (X'X)^{-1}X'E(Y)$ from part (a)

$= [(X'X)^{-1}][X'X]\beta$

$= I\beta = \beta$

(c) (4 points) Write the formula for the hat matrix $\mathbf{H}$ (that is, express it in terms of $\mathbf{X}$).

$H = X(X'X)^{-1}X'$

(d) (4 points) Show that the hat matrix $\mathbf{H}$ is idempotent.

$HH = X(X'X)^{-1}X'X(X'X)^{-1}X' = XX^{-1}(X')^{-1}X'X(X'X)^{-1}X' = IIX(X'X)^{-1}X' = H$

So $HH = H$, and H is idempotent.

(e) (5 points) Show that the variance-covariance matrix of the residuals, $Cov(\mathbf{e})$, where $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$, is equal to $\sigma^2(\mathbf{I} - \mathbf{H})$. Show each step explicitly.

We can begin with the estimator of our regression coefficients shown above

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

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

$\hat{Y} = HY$

So $e = Y - \hat{Y} = Y - HY = (I-H)Y$ ($I-H$: symmetric and idempotent)

Now when we apply

$Cov(e) = Cov[(I-H)Y]$

$= (I-H)Cov(Y)(I-H)'$

$= (I-H)Cov(Y)(I-H)$

$= (I-H)\sigma^2 I(I-H)$

$= \sigma^2(I-H)(I-H)$

$= \sigma^2(I-H)$