Access your exclusive
StudyPlus^{+} bonus content:
Flashcards | Career Guides | Samples
* Key Code Inside *

a/s/m
Exam SRM Study Manual

1st Edition, 2nd Printing

Abraham Weishaus, Ph.D., F.S.A., C.F.A., M.A.A.A.

NO RETURN IF OPENED

a/s/m
Actuarial Study Materials
Learning Made Easier

TO OUR READERS:

Please check A.S.M.'s web site at **www.studymanuals.com** for errata and updates. If you have any comments or reports of errata, please e-mail us at **mail@studymanuals.com**.

Contents

Lesson 2

Linear Regression: Estimating Parameters

Regression Modeling with Actuarial and Financial Applications 1.3, 2.1–2.2, 3.1–3.2; *An Introduction to Statistical Learning* 3.1–3.2, 3.3.2, 3.3.3

In a linear regression model, we have a variable y that we are trying to explain using variables $x_1, \ldots, x_k$.[1] We have n observations of sets of k explanatory variables and their responses: $\{y_i, x_{i1}, x_{i2}, \ldots, x_{ik}\}$ with $i = 1, \ldots, n$. We would like to relate y to the set of x_j, $j = 1, \ldots, k$ as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

where ε_i is an error term. We estimate the vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ by selecting the vector that minimizes $\sum_{i=1}^{n} \varepsilon_i^2$.

For statistical purposes, we make the following assumptions:

1. $\mathbf{E}[\varepsilon_i] = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. In other words, the variance of each error term is the same. This assumption is called *homoscedasticity* (sometimes spelled homoskedasticity).
2. ε_i are independent.
3. ε_i follow a normal distribution.

If these assumptions are valid, then for any set of values of the k variables $\{x_1, x_2, \ldots, x_k\}$, the resulting value of y will be normally distributed with mean $\beta_0 + \sum_{i=1}^{k} \beta_i x_i$ and variance σ^2. Moreover, the estimate of $\boldsymbol{\beta}$ is the maximum likelihood estimate.

Notice that our linear model has k parameters $\beta_1, \beta_2, \ldots, \beta_k$ in addition to the constant β_0. Thus we are really estimating $k + 1$ parameters. Some authors refer to "$k + 1$ variable regression". I've never been sure whether this is because $k + 1$ βs are estimated or because the response variable is counted as a variable.

2.1 Basic linear regression

When $k = 1$, the model is called "basic linear regression" or "simple linear regression".[2] In this case, the formulas for the estimators of β_0 and β_1 are

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{2.1}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.2}$$

Often we use Latin letters for the estimators of Greek parameters, so we can write b_i instead of $\hat{\beta}_i$.[3]

The formula for β_1 can be expressed as the quotient of the covariance of $\mathbf{x}$ and $\mathbf{y}$ over the variance of $\mathbf{x}$. The sample covariance is

$$cv_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

and the sample variance is

$$s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

[1] *Regression Modeling with Actuarial and Financial Applications* uses k for the number of variables, but *An Introduction to Statistical Learning* uses p.

[2] *Regression Modeling with Actuarial and Financial Applications* calls it basic linear regression and *An Introduction to Statistical Learning* calls it simple linear regression. As indicated in the previous paragraph, some authors call it "2 variable regression", and while this terminology is not used by either textbook, you may find it on old exam questions.

[3] *Regression Modeling with Actuarial and Financial Applications* uses b_i, while *An Introduction to Statistical Learning* uses $\hat{\beta}_i$.

The $n-1$s cancel when division is done, so they may be ignored. Then equation (2.1) becomes

$$\hat{\beta}_1 = \frac{cv_{xy}}{s_x^2}$$

You may use the usual shortcuts to calculate variance and covariance:

$$\text{Cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y]$$
$$\text{Var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

In the context of sample data, if we use the biased sample variance and covariance with division by n rather than $n-1$ (It doesn't really matter whether biased or unbiased is used, since the denominators of the sums, whether they are n or $n-1$, will cancel when one is divided by the other.), these formulas become

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\sum x_i \sum y_i}{n} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$
$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum x_i)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Let s_x, s_y, be the sample standard deviations of $\mathbf{x}$ and $\mathbf{y}$, and let r_{xy} be the sample correlation of $\mathbf{x}$ and $\mathbf{y}$, defined as follows:

$$r_{xy} = \frac{cv_{xy}}{s_x s_y}$$

From formula (2.1), we have $\hat{\beta}_1 = \frac{r_{xy} s_x s_y}{s_x^2}$, or

$$\hat{\beta}_1 = r_{xy}\frac{s_y}{s_x} \tag{2.3}$$

so $\hat{\beta}_1$ is proportional to the correlation of $\mathbf{x}$ and $\mathbf{y}$.

Example 2A You are given the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to fit to the following data:

x	2	2	3	4	5	5	7
y	34	38	38	53	50	60	70

Determine the least squares estimate of β_1.

Solution: First we calculate $\sum x_i^2$ and $\sum x_i y_i$, then we subtract $n\bar{x}^2$ and $n\bar{x}\bar{y}$. We obtain:

$$\sum x_i^2 = 132$$
$$\sum x_i y_i = 1510$$
$$\bar{x} = \frac{28}{7} = 4$$
$$\bar{y} = \frac{343}{7} = 49$$
$$\sum (x_i-\bar{x})^2 = 132 - 7(4^2) = 20$$
$$\sum (x_i-\bar{x})(y_i-\bar{y}) = 1510 - 7(4)(49) = 138$$
$$\hat{\beta}_1 = \frac{138}{20} = \boxed{\mathbf{6.9}}$$

Although not required by the question, we can easily calculate $\hat{\beta}_0$:

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\ &= 49 - (6.9)(4) = 21.4\end{aligned}$$ □

You would never go through the calculations of the previous example since your calculator can carry out the regression. On the TI-30XS, use `data`, ask for 2-Var statistics. In those statistics, item D is β_1 (with the unusual name `a`) and item E is β_0 (with the unusual name `b`). You can try this out on this quiz:

Quiz 2-1 For a new product released by your company, revenues for the first 4 months, in millions, are:

Month 1	27
Month 2	34
Month 3	48
Month 4	59

Revenues are assumed to follow a linear regression model of the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where x_i is the month and y_i is revenues.

Estimate β_1 for this model.

More likely, an exam question would give you summary statistics only and you'd use the formulas to get $\hat{\beta}_0$ and $\hat{\beta}_1$.

Example 2B For 8 observations of X and Y, you are given:

$$\bar{x} = 6 \qquad \bar{y} = 8 \qquad \sum x_i^2 = 408 \qquad \sum x_i y_i = 462$$

Perform a simple linear regression of Y on X:

$$y_i = \beta_0 + \beta_1 x_i$$

Determine $\hat{\beta}_0$.

Solution:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \\ &= \frac{462 - 8(6)(8)}{408 - 8(6^2)} = 0.65 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_0\bar{x} = 8 - 0.65(6) = \boxed{\mathbf{4.1}}\end{aligned}$$ □

The next example illustrates predicting an observation using the regression model.

Example 2C Experience for four cars on an automobile liability coverage is given in the following chart:

Miles Driven	7,000	10,000	11,000	12,000
Aggregate Claim Costs	600	2000	1000	1600

A least squares model relates aggregate claims costs to miles driven.

Calculate predicted aggregate claims costs for a car driven 5000 miles.

Solution: We let x_i be miles driven and y_i aggregate claim costs. It is convenient to drop thousands both in miles driven and aggregate claim costs.

$$\bar{x} = \frac{7+10+11+12}{4} = 10 \qquad \bar{y} = \frac{0.6+2+1+1.6}{4} = 1.3$$

$$\sum x_i^2 = 7^2+10^2+11^2+12^2 = 414 \qquad \sum x_i y_i = (7)(0.6)+(10)(2)+(11)(1)+(12)(1.6) = 54.4$$

$$\text{denominator} = 414-(4)(10^2) = 14 \qquad \text{numerator} = 54.4-(4)(10)(1.3) = 2.4$$

$$\hat{\beta}_1 = \frac{2.4}{14} = \frac{6}{35} \qquad \hat{\beta}_0 = 1300 - \left(\frac{6}{35}\right)(10000) = -\frac{2900}{7}$$

Notice that we multiplied back by 1000 when calculating $\hat{\beta}_0$.

The predicted value is therefore $-\frac{2900}{7}+\frac{6}{35}(5000) = \boxed{\mathbf{442.8571}}$. □

The fitted value of y_i, or $\hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}$, is denoted by $\hat{y}_i$. The difference between the actual and fitted values of y_i, or $\hat{\varepsilon}_i = y_i - \hat{y}_i$, is called the *residual*. As a result of the equations that are used to solve for $\hat{\beta}$, *the sum of the residuals* $\sum_{i=1}^{n} \hat{\varepsilon}_i$ *on the training set is always 0*. As with $\hat{\beta}_i$, we may use Latin letters instead of hats and denote the residual by e_i.

2.2 Multiple linear regression

Let's now discuss multiple regression, or $k > 1$. The generalized formulas involve matrices. We will use lower case boldface letters for column and row vectors and upper case boldface letters for matrices with more than one row and column. We will use a prime on a matrix to indicate its transpose. We define a column vector $\mathbf{x}_0$ whose values are all 1: $x_{10} = x_{20} = \cdots = x_{n0} = 1$. Then we can write $y_i = \sum_{j=0}^{k} \beta_j x_{ij}$, where β_0 is the coefficient of the column vector we just defined, rather than writing β_0 outside the sum. The matrix $\mathbf{X} = \{x_{ij}\}$ is an $n \times (k+1)$ matrix. The least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2.4}$$

and then the fitted value of y is $\hat{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$. I doubt you'd be expected to use formula (2.4) on an exam, unless you were given $(\mathbf{X}'\mathbf{X})^{-1}$, since it involves inverting a large matrix. In fact, I doubt you will be asked any matrix questions.

The $(\mathbf{X}'\mathbf{X})^{-1}$ matrix is singular (non-invertible) if there is a linear relationship among the column vectors of $\mathbf{X}$. Therefore, it is important that the column vectors not be collinear. Even if the variables are only "almost" collinear, the regression is unstable. We will discuss tests for collinearity in Section 5.3.

As with simple linear regression, the sum of the residuals is 0.

When an explanatory variable is a categorical variable with m possible values, you must include $m-1$ indicator variables in the model. Sometimes indicator variables are called "dummy variables". Each indicator variable corresponds to one possible value of the categorical variable. It is equal to 1 if the variable is equal to that value, 0 otherwise.

For example, if one of the explanatory variables is sex (male or female), you would set up one indicator variable for either male or female. If the indicator variable is for female, it would equal 0 if male and 1 if female. If one of the explanatory variables is age bracket and there are 5 age brackets, you would set up 4 indicator variables for 4 of the 5 age brackets. Notice that if you set up 5 variables, their sum would equal 1. The sum would be identical to $\mathbf{x_0}$, the first column vector of $\mathbf{X}$, resulting in a linear relationship among columns of the matrix, which would make it singular. Thus one variable must be omitted. The omitted variable is called the *base level* or *reference level*. You should select the value that occurs most commonly as the base level. If you select a value that is almost always 0, then the sum of the other indicator variables will almost always be 1, making the computation of the inverse of $\mathbf{X}'\mathbf{X}$ less stable.

A special case is a variable with only two categories. The indicator variable is then a binary variable.

2.3 Alternative model forms

Even though regression is a linear model, it is possible to incorporate nonlinear explanatory variables. Powers of variables may be included in the model. For example, you can estimate

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3 + \varepsilon_i$$

You can include interaction between explanatory variables by including a term multiplying them together:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

Another possibility is a regression with an exponential:

$$y_i = \beta_0 + \beta_1 e^{x_i} + \varepsilon_i$$

Linear regression assumes homoscedasticity, linearity, and normality. If these assumptions aren't satisfied, sometimes a few adjustments can be made to make the data satisfy these conditions.

Suppose the variance of the observations varies in a way that is known in advance. In other words, we know that $\text{Var}(\varepsilon_i) = \sigma^2/w_i$, with w_i varying by observation, although we don't necessarily know what σ^2 is. Then w_i is the *precision* of observation i, with $w_i = 0$ for an observation with no precision (which we would have to discard) and $w_i \to \infty$ for an exact observation. We can then multiply all the variables in observation i by $\sqrt{w_i}$. After this multiplication, all observations will have the same variance. Let $\mathbf{W}$ be the diagonal matrix with w_i in the i^{th} position on the diagonal, 0 elsewhere. Then equation (2.4) would be modified to

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} \tag{2.5}$$

The estimator $\hat{\boldsymbol{\beta}}^*$ is called the *weighted least squares estimator*.

One may also transform y to levelize the variance or to remove skewness. If variance appears to be proportional to y, logging y may levelize the variance:

$$\ln y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

which is equivalent to

$$y_i = e^{\beta_0 + \beta_1 x_i + \varepsilon_i}$$

In this model, $\ln y_i$ is assumed to have a normal distribution, which means that y_i is lognormal. A lognormal distribution is skewed to the right, so logging y may remove skewness.

A general family of power transformations is the *Box-Cox family of transformations*:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases} \tag{2.6}$$

This family includes taking y to any power, positive or negative, and logging. Adding a constant and dividing by a constant does not materially affect the form of a linear regression; it merely changes the intercept and scales the β coefficients. So $(y^\lambda - 1)/\lambda$ could just as well be y^λ. The only reason to subtract 1 and divided by λ is so that as $\lambda \to 0$, $(y^\lambda - 1)/\lambda \to \ln y$.

I doubt that the exam will require you to calculate parameters of regression models. Do a couple of the calculation exercises for this lesson just in case, but don't spend too much time on them.

Exercises

2.1. You are given the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to fit to the following data:

x	-2	-1	0	1	2
y	3	5	8	9	10

Determine the least squares estimate of β_0.

Exercises continue on the next page . . .

Table 2.1: Summary of Linear Model Formulas

For a simple regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.2}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{2.1}$$

$$\hat{\beta}_1 = r_{xy} \frac{s_y}{s_x} \tag{2.3}$$

For a multiple variable regression model

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2.4}$$

For any regression

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

For a weighted least squares model

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} \tag{2.5}$$

Box-Cox power transformations

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases} \tag{2.6}$$

2.2. You are fitting a linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to 18 observations.
You are given the following:

(i) $\sum_{i=1}^{18} x_i = 216$
(ii) $\sum_{i=1}^{18} x_i^2 = 3092$
(iii) $\sum_{i=1}^{18} y_i = 252$
(iv) $\sum_{i=1}^{18} y_i^2 = 4528$
(v) $\sum_{i=1}^{18} x_i y_i = 3364$

Determine the least squares estimate of β_1.

2.3. **[SRM Sample Question #17]** The regression model is $y = \beta_0 + \beta_1 x + \varepsilon$. There are six observations.
The summary statistics are:

$$\sum y_i = 8.5 \qquad \sum x_i = 6 \qquad \sum x_i^2 = 16 \qquad \sum x_i y_i = 15.5 \qquad \sum y_i^2 = 17.25$$

Calculate the least squares estimate of β_1.

(A) 0.1 (B) 0.3 (C) 0.5 (D) 0.7 (E) 0.9

Exercises continue on the next page . . .

2.4. **[SRM Sample Question #23]** Toby observes the following coffee prices in his company cafeteria:

- 12 ounces for 1.00
- 16 ounces for 1.20
- 20 ounces for 1.40

The cafeteria announces that they will begin to sell any amount of coffee for a price that is the value predicted by a simple linear regression using least squares of the current prices on size.

Toby and his co-worker Karen want to determine how much they would save each day, using the new pricing, if, instead of each buying a 24-ounce coffee, they bought a 48-ounce coffee and shared it.

Calculate the amount they would save.

(A) It would cost them 0.40 more.
(B) It would cost the same.
(C) They would save 0.40.
(D) They would save 0.80.
(E) They would save 1.20.

2.5. You are fitting the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to the following data:

x	2	5	8	11	13	15	16	18
y	−10	−9	−4	0	4	5	6	8

Determine the least squares estimate of β_1.

2.6. You are fitting the linear regression model $y_i = \beta_0 + \beta x_i + \varepsilon_i$ to the following data:

x	3	5	7	8	9	10
y	2	5	7	8	9	11

Determine the fitted value of y corresponding to $x = 6$.

2.7. You are fitting the linear regression model $\hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. You are given:

(i) $\sum_{i=1}^{28} x_i = 392$
(ii) $\sum_{i=1}^{28} y_i = 924$
(iii) $\sum_{i=1}^{28} x_i y_i = 13{,}272$
(iv) $\hat{\beta}_0 = -23$

Determine $\sum_{i=1}^{28} x_i^2$.

2.8. **[3-F84:5]** You are fitting the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to 10 points of data. You are given:

$$\sum x_i = 100$$
$$\sum y_i = 200$$
$$\sum x_i y_i = 2000$$
$$\sum x_i^2 = 2000$$
$$\sum y_i^2 = 5000$$

Calculate the least-squares estimate of β_1.

(A) 0.0 (B) 0.1 (C) 0.2 (D) 0.3 (E) 0.4

Exercises continue on the next page . . .

2.9. **[3L-S05:27]** Given the following information:

$$\sum x_i = 144$$
$$\sum y_i = 1{,}742$$
$$\sum x_i^2 = 2{,}300$$
$$\sum y_i^2 = 312{,}674$$
$$\sum x_i y_i = 26{,}696$$
$$n = 12$$

Determine the least squares equation for the following model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

(A) $\hat{y}_i = -0.73 + 12.16x_i$
(B) $\hat{y}_i = -8.81 + 12.16x_i$
(C) $\hat{y}_i = 283.87 + 10.13x_i$
(D) $\hat{y}_i = 10.13 + 12.16x_i$
(E) $\hat{y}_i = 23.66 + 10.13x_i$

2.10. **[120-F90:6]** You are estimating the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. You are given

i	1	2	3	4	5
x_i	6.8	7.0	7.1	7.2	7.4
y_i	0.8	1.2	0.9	0.9	1.5

Determine $\hat{\beta}_1$.

(A) 0.8 (B) 0.9 (C) 1.0 (D) 1.1 (E) 1.2

2.11. **[120-S90:11]** Which of the following are valid expressions for b, the slope coefficient in the simple linear regression of y on x?

I. $\dfrac{\left(\sum x_i y_i\right) - \bar{y}\sum x_i}{\left(\sum x_i^2\right) - \bar{x}\sum x_i}$

II. $\dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum x_i^2 - \bar{x}^2}$

III. $\dfrac{\sum x_i (y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

(A) I and II only (B) I and III only (C) II and III only (D) I, II and III
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

Exercises continue on the next page . . .

2.12. [**Old exam**] For the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon$ with 30 observations, you are given:

(i) $r_{xy} = 0.5$
(ii) $s_x = 7$
(iii) $s_y = 5$

where r_{xy} is the sample correlation coefficient.

Calculate the estimated value of β_1.

(A) 0.4 (B) 0.5 (C) 0.6 (D) 0.7 (E) 0.8

2.13. [**110-S83:14**] In a bivariate distribution the regression of the variable y on the variable x is $1500 + b(x - 68)$ for some constant b. If the correlation coefficient is 0.81 and if the standard deviations of y and x are 220 and 2.5 respectively, then what is the expected value of y, to the nearest unit, when x is 70?

(A) 1357 (B) 1515 (C) 1517 (D) 1643 (E) 1738

2.14. [**120-82-97:7**] You are given the following information about a simple regression model fit to 10 observations:

$$\sum x_i = 20$$
$$\sum y_i = 100$$
$$s_x = 2$$
$$s_y = 8$$

You are also given that the correlation coefficient $r_{xy} = -0.98$.

Determine the predicted value of y when $x = 5$.

(A) -10 (B) -2 (C) 11 (D) 30 (E) 37

2.15. In a simple regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, you are given

$$\sum x_i = 30 \qquad \sum y_i = 450$$
$$\sum x_i^2 = 270 \qquad \sum x_i y_i = 8100$$
$$n = 15$$
$$x_5 = 3 \qquad y_5 = 40$$

Calculate the fifth residual, $\hat{\varepsilon}_5$.

Exercises continue on the next page . . .

2.16. **[120-F89:13]** You are given:

Period	y	x_1	x_2
1	1.3	6	4.5
2	1.5	7	4.6
3	1.8	7	4.5
4	1.6	8	4.7
5	1.7	8	4.6

You are to use the following regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad i = 1, 2, \ldots 5$$

You have determined:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 1522.73 & 26.87 & -374.67 \\ 26.87 & 0.93 & -7.33 \\ -374.67 & -7.33 & 93.33 \end{pmatrix}$$

Calculate $\hat{\varepsilon}_2$.

(A) -0.2 (B) -0.1 (C) 0.0 (D) 0.1 (E) 0.2

2.17. You are fitting the following data to a linear regression model of the form $y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$:

y	5	3	10	4	3	5
x_1	0	1	0	1	0	1
x_2	1	0	1	1	0	1
x_3	0	1	1	0	0	0

You are given that

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{30}\begin{pmatrix} 26 & -10 & -18 & -12 \\ -10 & 20 & 0 & 0 \\ -18 & 0 & 24 & 6 \\ -12 & 0 & 6 & 24 \end{pmatrix}$$

Determine the least squares estimate of β_1.

Exercises continue on the next page . . .

2.18. **[120-82-94:11]** An automobile insurance company wants to use gender ($x_1 = 0$ if female, 1 if male) and traffic penalty points (x_2) to predict the number of claims (y). The observed values of these variables for a sample of six motorists are given by:

Motorist	x_1	x_2	y
1	0	0	1
2	0	1	0
3	0	2	2
4	1	0	1
5	1	1	3
6	1	2	5

You are to use the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad i = 1, 2, \ldots 6$$

You have determined

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{12}\begin{pmatrix} 7 & -4 & -3 \\ -4 & 8 & 0 \\ -3 & 0 & 3 \end{pmatrix}$$

Determine $\hat{\beta}_2$.

(A) -0.25 (B) 0.25 (C) 1.25 (D) 2.00 (E) 4.25

2.19. You are fitting the following data to the linear regression model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$:

y	1	2	6	5	1	2	3
x_1	0	0	1	-1	0	1	1
x_2	0	-1	0	0	1	-1	0
x_3	1	1	4	0	0	0	1

You are given that

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{30}\begin{pmatrix} 7 & 0 & 1.5 & -2.5 \\ 0 & 1.2 & 3 & -3 \\ 1.5 & 3 & 11.25 & -0.75 \\ -2.5 & -3 & -0.75 & 3.25 \end{pmatrix}.$$

Determine the fitted value of y for $x_1 = x_2 = x_3 = 1$.

Exercises continue on the next page . . .

2.20. [**Old exam**] You are examining the relationship between the number of fatal car accidents on a tollway each month and three other variables: precipitation, traffic volume, and the occurrence of a holiday weekend during the month. You are using the following model:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

y = the number of fatal car accidents
x_1 = precipitation, in inches
x_2 = traffic volumne
x_3 = 1, if a holiday weekend occurs during the month, and 0 otherwise

The following data were collected for a 12-month period:

Month	y	x_1	x_2	x_3
1	1	3	1	1
2	3	2	1	1
3	1	2	1	0
4	2	5	2	1
5	4	4	2	1
6	1	1	2	0
7	3	0	2	1
8	2	1	2	1
9	0	1	3	1
10	2	2	3	1
11	1	1	4	0
12	3	4	4	1

You have determined:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{6506}\begin{pmatrix} 257 & -82 & -446 \\ -82 & 254 & -364 \\ -446 & -364 & 2622 \end{pmatrix}$$

Determine $\hat{\beta}_1$.

(A) -0.07 (B) 0.15 (C) 0.24 (D) 0.70 (E) 1.30

2.21. [**S-F15:35**] You are given a regression model of liability claims with the following potential explanatory variables only:

- Vehicle price, which is a continuous variable modeled with a third order polynomial
- Average driver age, which is a continuous variable modeled with a first order polynomial
- Number of drivers, which is a categorical variable with four levels
- Gender, which is a categorical variable with two levels
- There is only one interaction in the model, which is between gender and average driver age.

Determine the maximum number of parameters in this model.

(A) Less than 9 (B) 9 (C) 10 (D) 11 (E) At least 12

Exercises continue on the next page . . .

2.22. [**MAS-I-S18:37**] You fit a linear model using the following two-level categorical variables:

$$X_1 = \begin{cases} 1 & \text{if Account} \\ 0 & \text{if Monoline} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if Multi-Car} \\ 0 & \text{if Single Car} \end{cases}$$

with the equation

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

This model produced the following parameter estimates:

$$\beta_0 = -0.10$$
$$\beta_1 = -0.25$$
$$\beta_2 = 0.58$$
$$\beta_3 = -0.20$$

Another actuary modeled the same underlying data, but coded the variables differently as such:

$$X_1 = \begin{cases} 0 & \text{if Account} \\ 1 & \text{if Monoline} \end{cases}$$

$$X_2 = \begin{cases} 0 & \text{if Multi-Car} \\ 1 & \text{if Single Car} \end{cases}$$

with the equation

$$E[Y] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2$$

Afterwards you make a comparison of the individual parameter estimates in the two models.

Calculate how many pairs of coefficient estimates $(\hat{\alpha}_i, \hat{\beta}_i)$ switched signs, and how many pairs of estimates stayed identically the same, when results of the two models are compared.

(A) 1 sign change, 0 identical estimates
(B) 1 sign change, 1 identical estimate
(C) 2 sign changes, 0 identical estimates
(D) 2 sign changes, 1 identical estimate
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

Solutions

2.1. $\bar{x} = 0$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, so $\hat{\beta}_0 = \bar{y} = \boxed{7}$.

2.2.

$$\sum (x_i - \bar{x})^2 = 3092 - \frac{216^2}{18} = 500$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 3364 - \frac{(216)(252)}{18} = 340$$

$$\hat{\beta}_1 = \frac{340}{500} = \boxed{\mathbf{0.68}}$$

2.3. The least squares estimate of β_1 is the covariance of x and y divided by the variance of x. In the following calculation, the numerator is n times the covariance and the denominator is n times the variance; the ns cancel. We have $n = 6$ observations.

$$b_1 = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 \Big/ n}$$
$$= \frac{15.5 - (6)(8.5)/6}{16 - 6^2/6} = \boxed{\mathbf{0.7}} \qquad \textbf{(D)}$$

2.4. The observations already lie in a straight line; each 4 ounce increase raises the price 0.20. The slope is therefore $0.2/4 = 0.05$ and the intercept (using 12 ounces = 1 = $0.05(12) + \beta_0$) is 0.4. By buying 48 ounces, one intercept, or $\boxed{\mathbf{0.40}}$, is saved. **(C)**

2.5. In the following, on the third line, because $\bar{y} = 0$, $\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum(x_i - \bar{x})y_i$.

$$\bar{x} = 11$$
$$\sum x_i^2 = 1188 \qquad \sum(x_i - \bar{x})^2 = 1188 - 8(11^2) = 220$$
$$\bar{y} = 0 \qquad \sum(x_i - \bar{x})(y_i - \bar{y}) = 2(-10) + 5(-9) + \cdots + 18(8) = 270$$
$$\hat{\beta}_1 = \frac{270}{220} = \boxed{\mathbf{1.2273}}$$

2.6.

$$\bar{x} = \bar{y} = 7$$
$$\sum(x_i - \bar{x})^2 = 4^2 + 2^2 + 0^2 + 1^2 + 2^2 + 3^2 = 34$$
$$\sum x_i y_i = (3)(2) + (5)(5) + (7)(7) + (8)(8) + (9)(9) + (10)(11) = 335$$
$$\sum(x_i - \bar{x})(y_i - \bar{y}) = 335 - 6(7)(7) = 41$$
$$\hat{\beta}_1 = \frac{41}{34}$$
$$\hat{\beta}_0 = 7 - \frac{41}{34}(7) = -\frac{49}{34}$$
$$\hat{y}(6) = -\frac{49}{34} + 6\left(\frac{41}{34}\right) = \frac{197}{34} = \boxed{\mathbf{5.7941}}$$

2.7.

$$\bar{y} = \frac{924}{28} = 33$$
$$\bar{x} = \frac{392}{28} = 14$$
$$\sum(x_i - \bar{x})(y_i - \bar{y}) = 13272 - 28(33)(14) = 336$$

$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, so

$$33 = -23 + \hat{\beta}_1(14)$$
$$\hat{\beta}_1 = 4 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$
$$\sum(x_i - \bar{x})^2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{4} = \frac{336}{4} = 84$$
$$\sum x_i^2 = \sum(x_i - \bar{x})^2 + 28\bar{x}^2 = 84 + 28(14^2) = \boxed{\mathbf{5572}}$$

2.8.

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{10}$$
$$= (2000) - \frac{(200)(100)}{10} = 0$$

It doesn't matter what the denominator $\sum(x_i - \bar{x})^2$ is; $\hat{\beta}_1 = \sum(x_i - \bar{x})(y_i - \bar{y}) \big/ \sum(x_i - \bar{x})^2 = \boxed{\mathbf{0}}$. **(A)**

2.9. By equation (2.1)

$$\beta_1 = \frac{12(26{,}696) - (144)(1{,}742)}{12(2{,}300) - 144^2} = \frac{69{,}504}{6{,}864} = \boxed{\mathbf{10.12587}}$$
$$\beta_0 = \frac{1{,}742}{12} - 10.12587\frac{144}{12} = 145.1667 - 10.12587(12) = \boxed{\mathbf{23.6562}} \qquad \textbf{(E)}$$

2.10.

$$\sum x_i = 35.5$$
$$\sum x_i^2 = 252.25$$
$$\sum(x_i - \bar{x})^2 = 252.25 - \frac{35.5^2}{5} = 0.2$$
$$\sum y_i = 5.3$$
$$\sum x_i y_i = 37.81$$
$$\sum(x_i - \bar{x})(y_i - \bar{y}) = 37.81 - \frac{(35.5)(5.3)}{5} = 0.18$$
$$\hat{\beta}_1 = \frac{0.18}{0.2} = \boxed{\mathbf{0.9}} \qquad \textbf{(B)}$$

2.11. The first one is correct, since it is equivalent to our formula

$$\frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

The second one is incorrect since $\bar{x}^2$ in the denominator should be multiplied by n.

The denominator of the third one is $\sum(x_i - \bar{x})^2$, like our formula. The numerator is the same as I, which is correct as we said above. **(B)**

2.12. Using equation (2.3),

$$\beta_1 = r_{xy}\frac{s_y}{s_x} = 0.5\left(\frac{5}{7}\right) = \boxed{\mathbf{0.3571}} \qquad \textbf{(A)}$$

2.13. Use equation (2.3).

$$b = r\frac{s_y}{s_x} = 0.81\left(\frac{220}{2.5}\right) = 71.28$$

The predicted value is $1500 + 71.28(70 - 68) = \boxed{\mathbf{1642.56}}$. **(D)**

2.14. Let the predicted value of y be y_5.

$$\hat{\beta}_1 = -0.98\left(\frac{8}{2}\right) = -3.92 \qquad \text{by equation (2.3)}$$

$$\hat{\beta}_0 = \bar{y} + 3.92\bar{x} = \frac{100}{10} + 3.92\left(\frac{20}{10}\right) = 17.84$$

$$y_5 = 17.84 - 3.92(5) = \boxed{\mathbf{-1.76}} \qquad \textbf{(B)}$$

2.15.

$$\hat{\beta}_1 = \frac{8100 - (30)(450)/15}{270 - 30^2/15} = 34\tfrac{2}{7}$$

$$\hat{\beta}_0 = \frac{450}{15} - 34\tfrac{2}{7}\left(\frac{30}{15}\right) = -38\tfrac{5}{7}$$

$$\hat{\varepsilon}_5 = 40 - \left(-38\tfrac{5}{7} + 34\tfrac{2}{7}(3)\right) = \boxed{\mathbf{-24\tfrac{2}{7}}}$$

2.16. We calculate $\hat{\boldsymbol{\beta}}$

$$\mathbf{X'y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 6 & 7 & 7 & 8 & 8 \\ 4.5 & 4.6 & 4.5 & 4.7 & 4.6 \end{pmatrix}\begin{pmatrix} 1.3 \\ 1.5 \\ 1.8 \\ 1.6 \\ 1.7 \end{pmatrix} = \begin{pmatrix} 7.9 \\ 57.3 \\ 36.19 \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y} = \begin{pmatrix} 9.9107 \\ 0.2893 \\ -2.2893 \end{pmatrix}$$

Then $\hat{\varepsilon}_2 = 1.5 - 9.9107 - 0.2893(7) + 2.2893(4.6) = \boxed{\mathbf{0.09498}}$. **(D)**

2.17.

$$\mathbf{X'Y} = \begin{pmatrix} 30 \\ 12 \\ 24 \\ 13 \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = \frac{1}{30}\begin{pmatrix} 72 \\ -60 \\ 114 \\ 96 \end{pmatrix}$$

$$\hat{\beta}_1 = \boxed{\mathbf{-2}}$$

2.18. The first coefficient of $\mathbf{X'y}$ is the sum of $\mathbf{y}$, or 12. The second is $1 + 3 + 5 = 9$ (not needed because $(\mathbf{X'X})^{-1}_{32} = 0$), and the third is $2(2) + 1(3) + 2(5) = 17$. Then

$$\hat{\beta}_2 = \frac{1}{12}\big((-3)(12) + 3(17)\big) = \frac{15}{12} = \boxed{\mathbf{1.25}}. \qquad \textbf{(C)}$$

2.19.

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 20 \\ 6 \\ -3 \\ 30 \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{30}\begin{pmatrix} 60.5 \\ -91.8 \\ -8.25 \\ 31.75 \end{pmatrix}$$

$$y(1,1,1) = \tfrac{1}{30}(60.5 - 91.8 - 8.25 + 31.75) = \boxed{\mathbf{-0.26}}$$

2.20. A little unusual not to have an intercept term β_0, but the formulas are the same as usual.
We need to compute $\mathbf{X}'\mathbf{y}$:

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 3 & 2 & 2 & 5 & 4 & 1 & 0 & 1 & 1 & 2 & 1 & 4 \\ 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 3 & 3 & 4 & 4 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 1 \\ 2 \\ 4 \\ 1 \\ 3 \\ 2 \\ 0 \\ 2 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 57 \\ 51 \\ 20 \end{pmatrix}$$

Then we multiply the first row of $(\mathbf{X}'\mathbf{X})^{-1}$ by $\mathbf{X}'\mathbf{y}$ to get the first coefficient of the β's, $\hat{\beta}_1$:

$$\frac{257(57) - 82(51) - 446(20)}{6506} = \frac{1547}{6506} = \boxed{\mathbf{0.2378}} \qquad \textbf{(C)}$$

2.21. A third order polynomial has 3 parameters that are multiplied by x, x^2, and x^3. A categorical variable with n levels has $n-1$ parameters. Thus there are 3 parameters for vehicle price, 1 for driver age, 3 for number of drivers, 1 for gender, and 1 for interaction. That sums up to 9. Add the intercept, and there are a total of $\boxed{\mathbf{10}}$ parameters. **(C)**

2.22. Since the model must produce the same results regardless of the values of the X_is, products of parameters and variables must be the same. Expressing the second model in terms of the first,

$$\begin{aligned} \mathbf{E}[Y] &= \alpha_0 + \alpha_1(1 - X_1) + \alpha_2(1 - X_2) + \alpha_3(1 - X_1)(1 - X_2) \\ &= \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 + (-\alpha_1 - \alpha_3)X_1 + (-\alpha_2 - \alpha_3)X_2 + \alpha_3 X_1 X_2 \end{aligned}$$

We see that $\alpha_3 = \beta_3$, but the relationships of the other parameters is not a simple sign change. **(E)**

Quiz Solutions

2-1.

$$\sum x_i = 1 + 2 + 3 + 4 = 10$$
$$\sum x_i^2 = 1^2 + 2^2 + 3^2 + 4^2 = 30$$
$$\sum y_i = 27 + 34 + 48 + 59 = 168$$
$$\sum x_i y_i = 27 + 2(34) + 3(48) + 4(59) = 475$$
$$\hat{\beta}_1 = \frac{475 - (10)(168)/4}{30 - 10^2/4} = \boxed{\mathbf{11}}$$

Practice Exam 1

1. A life insurance company is underwriting a potential insured as Preferred or Standard, for the purpose of determining the premium. Insureds with lower expected mortality rates are Preferred. The company will use factors such as credit rating, occupation, and blood pressure. The company constructs a decision tree, based on its past experience, to determine whether the potential insured is Preferred or Standard.

Determine, from a statistical learning perspective, which of the following describes this underwriting method.

I. Classification setting

II. Parametric

III. Supervised

(A) None (B) I and II only (C) I and III only (D) II and III only
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

2. An insurance company is modeling the probability of a claim using logistic regression. The explanatory variable is vehicle value. Vehicle value is banded, and the value of the variable is 1, 2, 3, 4, 5, or 6, depending on the band. Band 1 is the reference level.

The fitted value of the β corresponding to band 4 is -0.695.

Let O_1 be the odds of a claim for a policy in band 1, and O_4 the odds of a claim for a policy in band 4.

Determine O_4/O_1.

(A) 0.30 (B) 0.35 (C) 0.40 (D) 0.45 (E) 0.50

3. Auto liability claim size is modeled using a generalized linear model. Based on an analysis of the data, it is believed that the coefficient of variation of claim size is constant.

Which of the following response distributions would be most appropriate to use?

(A) Poisson (B) Normal (C) Gamma (D) Inverse Gamma (E) Inverse Gaussian

Exam questions continue on the next page . . .

4. You are given the following output from a GLM to estimate loss size:

(i) Distribution selected is Inverse Gaussian.
(ii) The link is $g(\mu) = 1/\mu^2$.

Parameter	β
Intercept	0.00279
Vehicle Body	
Coupe	0.002
Sedan	−0.001
SUV	0.003
Vehicle Value (000)	−0.00007
Area	
B	−0.025
C	0.015
D	0.005

Calculate mean loss size for a sedan with value 25,000 from Area A.

(A) Less than 80
(B) At least 80, but less than 160
(C) At least 160, but less than 320
(D) At least 320, but less than 640
(E) At least 640

1280

5. For a generalized linear model,

(i) There are 72 observations.
(ii) There are 25 parameters.
(iii) The loglikelihood is −361.24

You are considering adding a cubic polynomial variable.

Determine the lowest loglikelihood for which this additional variable would be accepted at 1% significance.

(A) −358 (B) −356 (C) −354 (D) −352 (E) −350

6. In a principal components analysis, there are 2 variables. The loading of the first principal component on the first variable is −0.6 and the loading of the first principal component on the second variable is positive. The variables have been centered at 0.

For the observation $(0.4, x_2)$, the first principal component score is 0.12.

Determine x_2.

(A) 0.25 (B) 0.30 (C) 0.35 (D) 0.40 (E) 0.45

Exam questions continue on the next page . . .

7. Determine which of the following statements is/are true.

I. The lasso is a more flexible approach than linear regression.

II. Flexible approaches lead to more accurate predictions.

III. Generally, more flexible approaches result in less bias.

(A) I only (B) II only (C) III only (D) I, II, and III
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

8. A generalized linear model for automobile insurance with 40 observations has the following explanatory variables:

SEX (male or female)
AGE (4 levels)
TYPE OF VEHICLE (sedan, coupe, SUV, van)
MILES DRIVEN (continuous variable)
USE (business, pleasure, farm)

Model I includes all of these variables and an intercept. Model II is the same as Model I except that it excludes USE. You have the following statistics from these models:

	Deviance	AIC
Model I	23.12	58.81
Model II		62.61

Using the likelihood ratio test, which of the following statements is correct?

(A) Accept USE at 0.5% significance.
(B) Accept USE at 1.0% significance but not at 0.5% significance.
(C) Accept USE at 2.5% significance but not at 1.0% significance.
(D) Accept USE at 5.0% significance but not at 2.5% significance.
(E) Reject USE at 5.0% significance.

9. You are given the following two clusters:

{(8,2), (9,7), (12,5)} and {(10,3), (11,1)}

Calculate the dissimilarity measure between the clusters using Euclidean distance and average linkage.

(A) 3.6 (B) 3.7 (C) 3.8 (D) 3.9 (E) 4.0

10. A normal linear model with 2 variables and an intercept is based on 45 observations. $\hat{y}_j$ is the fitted value of y_j, and $\hat{y}_{j(i)}$ is the fitted value of y_j if observation i is removed. You are given:

(i) $\sum_{j=1}^{45}(\hat{y}_j - \hat{y}_{j(1)})^2 = 4.1$.
(ii) The leverage of the first observation is 0.15.

Determine $|\hat{\varepsilon}_1|$, the absolute value of the first residual of the regression with no observation removed.

(A) 3.9 (B) 4.4 (C) 4.9 (D) 5.4 (E) 5.9

Exam questions continue on the next page . . .

11. A least squares model with a large number of predictors is fitted to 90 observations. To reduce the number of predictors, forward stepwise selection is performed.

For a model with k predictors, RSS = c_k.

The estimated variance of the error of the fit is $\hat{\sigma}^2 = 40$.

Determine the value of $c_d - c_{d+1}$ for which you would be indifferent between the $d+1$-predictor model and the d-predictor model based on Mallow's C_p.

(A) 40 (B) 50 (C) 60 (D) 70 (E) 80

12. A classification response variable has three possible values: A, B, and C.

A split of a node with 100 observations in a classification tree resulted in the following two groups:

Group	Number of A	Number of B	Number of C
I	40	10	10
II	5	25	10

Calculate the cross-entropy for this split.

(A) 0.72 (B) 0.76 (C) 0.80 (D) 0.84 (E) 0.88

13. Determine which of the following statements are true regarding cost complexity pruning.

I. A higher α corresponds to higher MSE for the training data.

II. A higher α corresponds to higher bias for the test data.

III. A higher α corresponds to a higher $|T|$.

(A) None (B) I and II only (C) I and III only (D) II and III only
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

14. Determine which of the following constitutes data snooping.

(A) Using personal data without authorization of the individuals.
(B) Using large amounts of low-quality data.
(C) Using an excessive number of variables to fit a model.
(D) Fitting an excessive number of models to one set of data.
(E) Validating a model with a large number of validation sets.

15. Determine which of the following statements are true regarding K-nearest neighbors (KNN) regression.

I. KNN tends to perform better as the number of predictors increases.

II. KNN is easier to interpret than linear regression.

III. KNN becomes more flexible as $1/K$ increases.

(A) None (B) I and II only (C) I and III only (D) II and III only
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

Exam questions continue on the next page . . .

16. A department store is conducting a cluster analysis to help focus its marketing. The store sells many different products, including food, clothing, furniture, and computers. Management would like the clusters to group together customers with similar shopping patterns.

Determine which of the following statements regarding cluster analysis for this department store is/are true.

I. The clusters will depend on whether the input data is units sold or dollar amounts sold.

II. Hierarchical clustering would be preferable to K-means clustering.

III. If a correlation-based dissimilarity measure is used, frequent and infrequent shoppers will be grouped together.

(A) I only (B) II only (C) III only (D) I, II, and III
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

17. Determine which of the following statements regarding principal components analysis is/are true.

I. Principal components analysis is a method to visualize data.

II. Principal components are in the direction in which the data is most variable.

III. Principal components are orthogonal.

(A) I only (B) II only (C) III only (D) I, II, and III
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

18. A random walk is the cumulative sum of a white noise process c_t. You are given that c_t is normally distributed with mean 0 and variance σ^2.

Which of the following statements are true?

I. The mean of the random walk does not vary with time.

II. At time 50, the variance is $50\sigma^2$.

III. Differences of the random walk form a stationary time series.

(A) I only (B) II only (C) III only (D) I, II, and III
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

19. You are given the following regression model, based on 22 observations.

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4 + \beta_5\mathbf{x}_5 + \varepsilon$$

The residual sum of squares for this model is 156.

If the variables $\mathbf{x}_4$ and $\mathbf{x}_5$ are removed, the error sum of squares is 310.

Calculate the F ratio to determine the significance of the variables $\mathbf{x}_4$ and $\mathbf{x}_5$.

(A) 3.9 (B) 4.9 (C) 5.9 (D) 6.9 (E) 7.9

20. You are given the following time series:

20, 22, 21, 24, 23

The time series is fitted to an AR(1) process with $y_t = 20.325 + 0.1y_{t-1}$.

Calculate the estimated variance of the residuals.

(A) 1.3 (B) 1.7 (C) 2.1 (D) 2.5 (E) 2.9

Exam questions continue on the next page . . .

21. Determine which of the following algorithms is greedy.

I. Hierarchical clustering algorithm

II. Recursive binary splitting algorithm for decision trees

III. Forward subset selection algorithm

(A) I only (B) II only (C) III only (D) I, II, and III
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

22. Determine which of the following statements about boosting is/are true.

I. Selecting B too high can result in overfitting.

II. Selecting a low shrinkage parameter tends to lead to selecting a lower B.

III. If $d = 1$, the model is an additive model.

(A) None (B) I and II only (C) I and III only (D) II and III only
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

23. To validate a time series model based on 20 observations, the first 15 observations were used as a model development subset and the remaining 5 observations were used as a validation subset. The actual and fitted values for those 5 observations are

t	y_t	$\hat{y}_t$
16	7	10
17	9	12
18	12	14
19	18	16
20	22	18

Calculate the MSE.

(A) 7.4 (B) 8.4 (C) 9.5 (D) 10.5 (E) 11.5

24. In a hurdle model, the probability of overcoming the hurdle is 0.7. If the hurdle is overcome, the count distribution is $kg(j)$, where $g(j)$ is the probability function of a Poisson distribution with parameter $\lambda = 0.6$.

Calculate the probability of 1.

(A) 0.23 (B) 0.31 (C) 0.39 (D) 0.45 (E) 0.51

25. For a generalized linear model, you are given

(i) The negative loglikelihood of the model is 74.88.
(ii) The deviance of the model is 8.70.
(iii) The maximized loglikelihood of the minimal model is −90.31.

Calculate the pseudo-R^2 statistic.

(A) 0.64 (B) 0.68 (C) 0.71 (D) 0.74 (E) 0.78

Exam questions continue on the next page . . .

26. The number of policies sold by an agent in a year, y, is modeled as a function of the number of years of experience, x. The model is a Poisson regression with a log link. The fitted coefficient of x is $\beta_1 = 0.06$.

The expected number of policies sold after 2 years of experience is a and the expected number of policies sold after 5 years of experience is b.

Calculate b/a.

(A) 1.18 (B) 1.19 (C) 1.20 (D) 1.21 (E) 1.22

27. Which of the following statements are true?

I. Partial Least Squares is a supervised method of dimension reduction.

II. Partial Least Squares directions are linear combinations of the original variables.

III. Partial Least Squares can be used for feature selection.

(A) None (B) I and II only (C) I and III only (D) II and III only
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

28. Disability income claims are modeled using linear regression. The model has two explanatory variables:

1. *Occupational class*. This may be (1) professional with rare exposure to hazards, (2) professional with some exposure to hazards, (3) light manual labor, (4) heavy manual labor.
2. *Health*. This may be (1) excellent, (2) good, (3) fair.

The model includes an intercept and all possible interactions.

Determine the number of interaction parameters β_i in the model.

(A) 6 (B) 8 (C) 9 (D) 11 (E) 12

29. Consider the vector $\{5, -3, 8, -2, 4\}$.

Calculate the absolute difference between the ℓ_2 norm and ℓ_1 norm of this vector.

(A) 11 (B) 13 (C) 15 (D) 17 (E) 19

30. For a simple linear regression of **y** on **x**:

(i) There are 25 observations.
(ii) $\bar{x} = 32$
(iii) The unbiased sample variance of x is 20.
(iv) $x_4 = 22$

Calculate the leverage of x_4.

(A) 0.21 (B) 0.23 (C) 0.25 (D) 0.27 (E) 0.29

31. You are given the time series

182, 138, 150, 192, 177

The series is smoothed using exponential smoothing with $w = 0.8$.

Calculate the sum of squared one-step prediction errors.

(A) 2042 (B) 2555 (C) 3038 (D) 3589 (E) 3966

Exam questions continue on the next page . . .

32. Determine which of the following statements about classification trees is/are true.

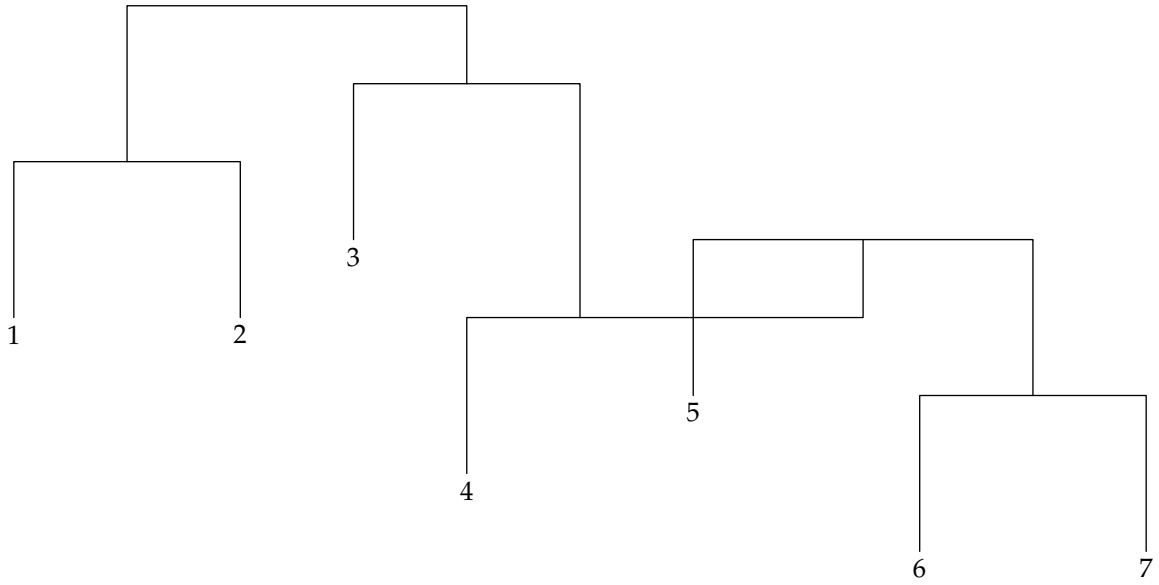
I. Classification error is not sensitive enough for growing trees.

II. Classification error is not sensitive enough for pruning trees.

III. The predicted values of two terminal nodes coming out of a split are different.

(A) I only (B) II only (C) III only (D) I, II, and III
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

33. Hierarchical clustering is performed on 7 observations, resulting in the following dendrogram:



Determine which of the following statements is/are true.

I. Centroid linkage was used.

II. Observation 3 is closer to observation 4 than to observation 7.

III. Observations 3 and 4 are closer to each other than observations 1 and 2.

(A) I only (B) II only (C) III only (D) I, II, and III
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

34. For a simple linear regression of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, you are given

(i) $\bar{y} = 100$
(ii) $\sum_{i=1}^{8} y_i^2 = 81{,}004$
(iii) $\sum_{i=1}^{8} \hat{y}_i^2 = 80{,}525$

Calculate R^2.

(A) 0.46 (B) 0.48 (C) 0.50 (D) 0.52 (E) 0.54

Exam questions continue on the next page . . .

35. Determine which of the following are results of overfitting models.

I. The residual standard error may increase.

II. The model may be more difficult to interpret.

III. The variables may be collinear.

(A) I only (B) II only (C) III only (D) I, II, and III
(E) The correct answer is not given by (A) , (B) , (C) , or (D) .

Solutions to the above questions begin on page 353.

Appendix A. Solutions to the Practice Exams

Answer Key for Practice Exam 1

1	C	11	E	21	E	31	C
2	E	12	E	22	C	32	A
3	C	13	B	23	B	33	A
4	B	14	D	24	E	34	D
5	B	15	E	25	E	35	D
6	E	16	D	26	C		
7	C	17	D	27	B		
8	C	18	D	28	A		
9	C	19	E	29	A		
10	B	20	D	30	C		

Practice Exam 1

1. **[Lesson 1]** Classification setting—the company is choosing a class. Supervised—there is something being predicted. But decision trees are not parametric. **(C)**

2. **[Lesson 12]** In logistic regression, $g(\mu)$ is the logarithm of the odds, so we must exponentiate β to obtain odds ratio.

$$e^{-0.695} = \boxed{\mathbf{0.4991}} \qquad \textbf{(E)}$$

3. **[Section 11.1]** The square of the coefficient of variation is the variance divided by the square of the mean. If it is constant, then variance is proportional to mean squared. This is true for a gamma distribution. **(C)**

4. **[Section 11.1]** Area A is the base level, so nothing is added to $g(\mu)$ for it.

$$g(\mu) = 0.00279 - 0.001 + 25(-0.00007) = 0.00004$$
$$\frac{1}{\mu^2} = 0.00004$$
$$\mu = \sqrt{\frac{1}{0.00004}} = \boxed{\mathbf{158.11}} \qquad \textbf{(B)}$$

5. **[Section 14.2]** A cubic polynomial adds 3 parameters. The 99$^{\text{th}}$ percentile of chi-square at 3 degrees of freedom is 11.345. Twice the difference in loglikelihoods must exceed 11.345, so the loglikelihood must increase by 5.67. Then $-361.24 + 5.67 = \boxed{\mathbf{-355.57}}$. **(B)**

6. **[Section 17.1]** The loading of the first principal component on the second variable is $\sqrt{1 - 0.6^2} = 0.8$. We are given

$$-0.6(0.4) + 0.8x_2 = 0.12$$

It follows that $x_2 = \boxed{\mathbf{0.45}}$. **(E)**

7. **[Lesson 1]**

I. The lasso is more restrictive than linear regression.✗

II. Flexible approaches may not lead to more accurate predictions due to overfitting.✗

III. This sentence is lifted from *An Introduction to Statistical Learning* page 35.✓

(C)

8. **[Lesson 14]** USE has 3 levels, so Model II has 2 parameters fewer than Model I. Thus the AIC penalty on Model II is 4 less than for Model I. The AIC for Model I is 3.80 less than for Model II, but before the penalty, twice the negative loglikelihood of Model I is 7.80 less than for Model II. The critical values for chi-square with 2 degrees of freedom are 7.378 at 2.5% and 9.210 at 1%, making **(C)** the correct answer choice.

9. **[Section 18.2]** We have to calculate all 6 distances between points and average them.

Point 1	Point 2	Distance
(8,2)	(10,3)	$\sqrt{5}$
(9,7)	(10,3)	$\sqrt{17}$
(12,5)	(10,3)	$\sqrt{8}$
(8,2)	(11,1)	$\sqrt{10}$
(9,7)	(11,1)	$\sqrt{40}$
(12,5)	(11,1)	$\sqrt{17}$

The average distance is $1/6\left(\sqrt{5}+\sqrt{17}+\sqrt{8}+\sqrt{10}+\sqrt{40}+\sqrt{17}\right) = \boxed{\mathbf{3.7996}}$. **(C)**

10. **[Section 5.2]** Use the second equality of formula (5.2). The standard error of the first residual is $s\sqrt{1-h_{11}}$.

$$\frac{4.1}{3s^2} = \left(\frac{\hat{\varepsilon}_1}{s\sqrt{1-0.15}}\right)^2\left(\frac{0.15}{3(0.85)}\right)$$

$$4.1 = \left(\frac{\hat{\varepsilon}_1^2}{0.85}\right)\left(\frac{0.15}{0.85}\right)$$

$$\hat{\varepsilon}_1^2 = \frac{4.1(0.85^2)}{0.15} = 19.7483$$

$$|\hat{\varepsilon}_1| = \boxed{\mathbf{4.4439}} \qquad \textbf{(B)}$$

11. **[Section 7.2]** We will use the definition of Mallow's C_p from *An Introduction to Statistical Learning*, but you would get the same result using the definition in *Regression Modeling with Actuarial and Financial Applications*.

$C_p = \frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2)$, and we can ignore $1/n$. So we want

$$c_d + 2d(40) = c_{d+1} + 2(d+1)(40)$$

This implies

$$c_d - c_{d+1} = 2(40) = \boxed{\mathbf{80}} \qquad \textbf{(E)}$$

12. **[Section 16.1]** We weight the cross-entropies for the two groups with the proportions of observations in each group, 0.6 and 0.4

$$D = -0.6\left(\frac{2}{3}\ln\frac{2}{3} + \frac{1}{6}\ln\frac{1}{6} + \frac{1}{6}\ln\frac{1}{6}\right) - 0.4\left(\frac{1}{8}\ln\frac{1}{8} + \frac{5}{8}\ln\frac{5}{8} + \frac{1}{4}\ln\frac{1}{4}\right) = \boxed{\mathbf{0.88064}} \qquad \textbf{(E)}$$

13. **[Section 16.1]** Higher α means more tree pruning and fewer nodes. That will increase the MSE on the training data and raise bias on the test data. $|T|$ is the number of terminal nodes, which decreases. **(B)**

14. **[Section 7.1]** Data snooping refers to **(D)**.

15. **[Lesson 15]**

I. KNN tends to perform worse as the number of predictors increases, since the points tend to be further apart. ✗

II. Linear regression is easier to interpret than KNN. ✗

III. KNN is most flexible as K get smaller, or as $1/K$ increases. ✓

(E)

16. **[Section 18.2]**

I. Furniture sales have low units but high dollar amounts, and food is the other way around, so the input data would have quite different patterns, with significant effect on clusters.✓

II. A correlation-based dissimilarity method is desirable, and that is much easier to use with hierarchical clustering.✓

III. Correlation is scale-free, so frequent and infrequent shoppers with the same shopping patterns would be grouped together.✓

(D)

17. **[Section 17.1]** All three statements are true. **(D)**

18. **[Lesson 19]** All three statements are true. The mean of c_t is 0 so the mean of sums of c_t is also 0, a constant. The variance at time t is $t\sigma^2$, here $50\sigma^2$. Differences of the series are white noise, which is stationary. **(D)**

19. **[Section 4]** There are $n = 22$ observations, $k + 1 = 6$ coefficients in the unrestricted model, and $q = 2$ restrictions.

$$F_{2,16} = \frac{(\text{Error SS}_R - \text{Error SS}_{UR})/q}{\text{Error SS}_{UR}/(n-k-1)} = \frac{(310-156)/2}{156/16} = \boxed{\textbf{7.897}} \qquad \textbf{(E)}$$

20. **[Lesson 20]** The residuals are

$$\begin{aligned} 22 - \big(20.325 + 0.1(20)\big) &= -0.325 \\ 21 - \big(20.325 + 0.1(22)\big) &= -1.525 \\ 24 - \big(20.325 + 0.1(21)\big) &= 1.575 \\ 23 - \big(20.325 + 0.1(24)\big) &= 0.275 \end{aligned}$$

The mean of the residuals is 0. The estimated variance of the residuals, by formula (20.3), is

$$s^2 = \frac{1}{2}\big((-0.325)^2 + (-1.525)^2 + 1.575^2 + 0.275^2\big) = \boxed{\textbf{2.49375}} \qquad \textbf{(D)}$$

21. **[Section 7.1, Lessons 16, and Section 18.2]** II and III are greedy in that they select the best choice at each step and don't consider later steps. While hierarchical clustering selects the least dissimilar cells at each iteration, there is no particular measure that would indicate whether a better clustering is possible with a different choice, so it is not considered greedy. **(E)**

22. **[Section 16.2]** I and III are true. The opposite of II is true: a low shrinkage parameter leads to selecting a higher B since less is learned at each iteration, so more time is needed to learn **(C)**

23. **[Section 19.6]** MSE is the mean square error, with division by 5 rather than 4, since the fit is not a function of the validation subset. The residuals are $-3, -3, -2, 2, 4$

$$\text{MSE} = \frac{3^2 + 3^2 + 2^2 + 2^2 + 4^2}{5} = \boxed{8.4} \qquad \textbf{(B)}$$

24. **[Subsection 13.3.2]** k is the quotient $(1-\pi)/(1-g(0))$, where π is the probability of 0 (0.3 here) and $g(0)$ is the Poisson probability of 0, which is $e^{-0.6}$ here. The probability of 1 is

$$p_1 = \left(\frac{1-0.3}{1-e^{-0.6}}\right)0.6e^{-0.6} = \boxed{0.510875} \qquad \textbf{(E)}$$

25. **[Section 14.5]** The deviance is twice the excess of the loglikelihood of the saturated model, $l_{\max}$, over the loglikelihood of the model under consideration, $l(\mathbf{b})$, so

$$\begin{aligned} 2\big(l_{\max} - l(\mathbf{b})\big) &= 8.70 \\ l_{\max} + 74.88 &= 4.35 \\ l_{\max} &= -70.53 \end{aligned}$$

The pseudo-R^2 statistic is

$$\text{pseudo-}R^2 = \frac{l(\mathbf{b}) - l_0}{l_{\max} - l_0} = \frac{-74.88 + 90.31}{-70.53 + 90.31} = \boxed{0.78} \qquad \textbf{(E)}$$

26. **[Section 13.1]** In a Poisson regression with a log link, the ratio of expected values is the exponential of the difference of the xs. Here, that is $e^{0.06(5-2)} = \boxed{1.1972}$. **(C)**

27. **[Section 8.2]**

1. PLS is a supervised method since it takes the response into account when determining the coefficients. ✓
2. In both dimension reduction methods we study, the selected directions are linear combinations of the original variables. ✓
3. PLS creates new variables that are functions of the original ones, so it does not select features. ✗

(B)

28. **[Lesson 2]** For each explanatory variable there is a base level. There are 3 non-base occupational classes and 2 non-base health classes. Thus there are $3 \times 2 = \boxed{6}$ interaction parameters. **(A)**

29. **[Section 8.1]** Let v be the vector.

$$\begin{aligned} \|v\|_1 &= 5 + 3 + 8 + 2 + 4 = 22 \\ \|v\|_2 &= \sqrt{5^2 + 3^2 + 8^2 + 2^2 + 4^2} = 10.8628 \end{aligned}$$

The absolute difference is $|22 - 10.8628| = \boxed{11.1372}$. **(A)**

30. **[Section 5.2]** Use formula (5.1):

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

The denominator is $n - 1$ times the sample variance. We get

$$h_{44} = \frac{1}{25} + \frac{(22-32)^2}{20(24)} = \boxed{\mathbf{0.24833}} \qquad \textbf{(C)}$$

31. **[Section 21.2]** The predictions are

$$x_{2|1} = 182$$
$$x_{3|2} = 0.2(138) + 0.8(182) = 173.2$$
$$x_{4|3} = 0.2(150) + 0.8(173.2) = 168.56$$
$$x_{5|4} = 0.2(192) + 0.8(168.56) = 173.248$$

The sum of squared errors is $(-44)^2 + (-23.2)^2 + 23.44^2 + (-3.752)^2 = \boxed{\mathbf{3037.751}}$. **(C)**

32. **[Section 16.1]** I is true. But classification error is preferred for pruning tree since that is the measure of predictive accuracy. And the predicted values of two terminal nodes coming out of a split may be the same, due to different levels of node purity. **(A)**

33. **[Section 18.2]**

I. There is an inversion; the split between {4} and {5,6,7} is at a lower level than the split between {5} and {6,7}, and of the four linkages we studied, only centroid has inversions.✓

II. {3} is fused with {4,5,6,7}, so it is no closer to {4} than to {7}.✗

III. {1} and {2} fuse at a lower level than {3} and {4,5,6,7}, so {1} and {2} are closer.✗

(A)

34. **[Section 3.2]**

$$\text{Total SS} = 81{,}004 - 8(100^2) = 1{,}004$$
$$\text{Regression SS} = 80{,}525 - 8(100^2) = 525$$
$$R^2 = \frac{525}{1004} = \boxed{\mathbf{0.52}} \qquad \textbf{(D)}$$

35. **[Lesson 10]** All three statements are true. **(D)**