Linear Statistical Models

Inference Topics Covered So Far

- Identified estimators for common parameters.
- Discussed the sampling distributions of estimators.
- Introduced ways to judge the "goodness" of an estimator. (bias, MSE, etc.)
- Used maximum likelihood estimation.
- Used confidence intervals and hypothesis testing to make inferences about means and proportions.

None of the methods that we've discussed so far allow us to model the relationship (correlation) between two variables.

Describing the Linear Relationship

- Imagine that you have two quantitative variables that are correlated.
- Think back to our coefficient of correlation, ρ. Now instead of just measuring the strength of the linear relationship, we want to get a more specific idea about the nature of the relationship.
- In particular, we might want to predict values of one variable given the other variable.
- We will discuss how to make inference when two variables have linear relationship.
- If the relationship between the 2 variables is not linear, sometimes appropriate transformations of the data may yield a more linear relationship.

Deterministic Linear Relationships

- If the correlation between the two variables is known to be perfect (1 or -1), or very close to it, we might use a deterministic model.
- Linear relationship between two variables can be described based on the equation for a line.
- Then, if we know the value of one variable, we can exactly (or very close to it) predict the value of the other variable.
- This kind of models may be appropriate for well-established laws of science.
- One can model this kind of linear relationship between two variables X and Y as $Y = \beta_0 + \beta_1 X$ where β_0 and β_1 are the intercept and slope of the line that describes the relationship.

Probabilistic Linear Models

What if the correlation between two variables is not perfect? What if there seems to be a scatter cloud of points that has a general linear trend?

- We can express the relationship using the model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- For every observation (X_i, Y_i), Y_i is a linear function of X_i plus some random "noise" given by ε_i.
- The noise is assumed to have mean 0 and be independent from data point to data point.
- We cannot exactly predict Y for every X, but we can say what the expected value for Y given X is. Then, we are modeling the means of the Y_is given the changing X_is: $E(Y_i) = \beta_0 + \beta_1 X_i$

- This is called the regression of Y on X.
- In regression, we have a dependent and one or more independent variable. The role of dependent variable is different from that of independent variables.
- One important goal of regression is to predict the value of a variable form the value of the other variables. The variable that has to be predicted is chosen as the dependent variable and the predictor variables are chosen as the independent variables.
- When we have only one independent variable, then that is called simple linear regression. When we have more than one independent variable, that is called multiple linear regression.
- Above is a simple linear regression model with X as the independent and Y as the dependent variable.

Example of Two Correlated Variables

We may be interested in predicting a student's second midterm score, given his/her first midterm score. We could look at a sample of such scores to help us determine how these grades are correlated.

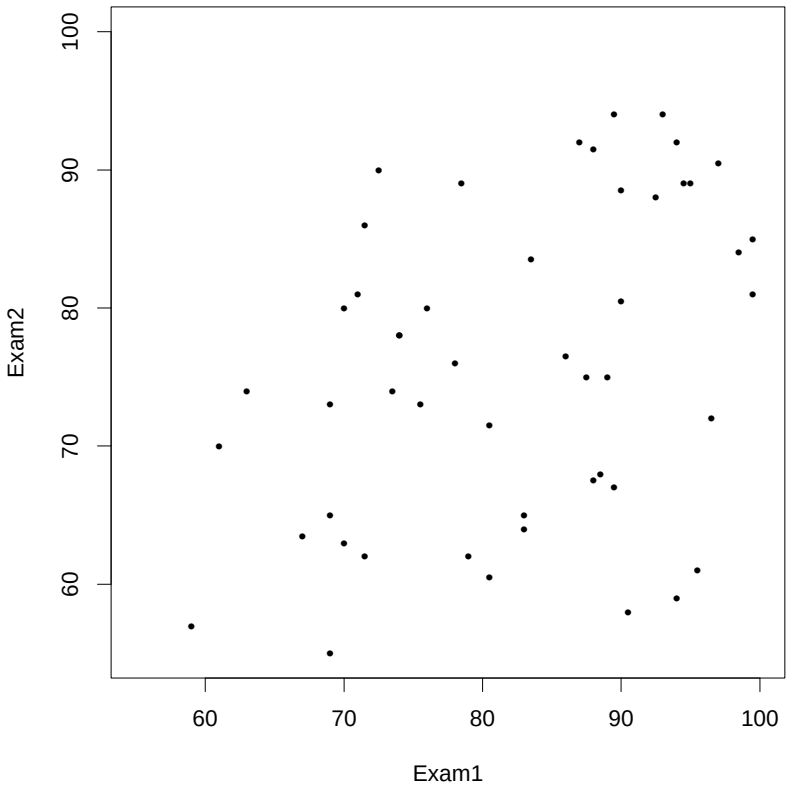

Figure 1: Sample of scores on the two midterms

Interpretations of Estimates

- Since we have just a sample from the population, we can only estimate the slope and intercept of the "true" regression line.
- These estimates are denoted $\hat{\beta}_0$ and $\hat{\beta}_1$, and the fitted values that they yield are denoted $\hat{y}$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- $\hat{\beta}_1$ gives the estimated change in the dependent variable associated with a one unit change in the independent variable.
- $\hat{\beta}_0$ gives the estimated value of the dependent variable when independent variable is 0.

How to Fit the Line?

- We want the line to be as close to the data points as possible, but since there is so much variation from a strict linear pattern, we need some criterion to choose the "best" line.
- How do we measure the distances between the data points and the line? We could use vertical distance, horizontal distance, or closest distance (perpendicular approach from each data point to the fitted line).
- The vertical distance between the data point and the fitted line is appropriate because that is an estimate for the "noise" ε_i.

Least Squares Approach

- The vertical distance between a data point y_i and the regression line is called the error or residual, and denoted by $e_i = y_i - \hat{y}_i$. The sum of squared errors (SSE) is $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.
- We choose the line that minimizes the SSE.
- The approach is called "least squares" since it minimizes the sum of the squared vertical distances. The sum of the vertical distances, not squared, is 0.
- This provides us with the following least square estimates for the slope and intercept of the regression line:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

Back to Our Grades Data

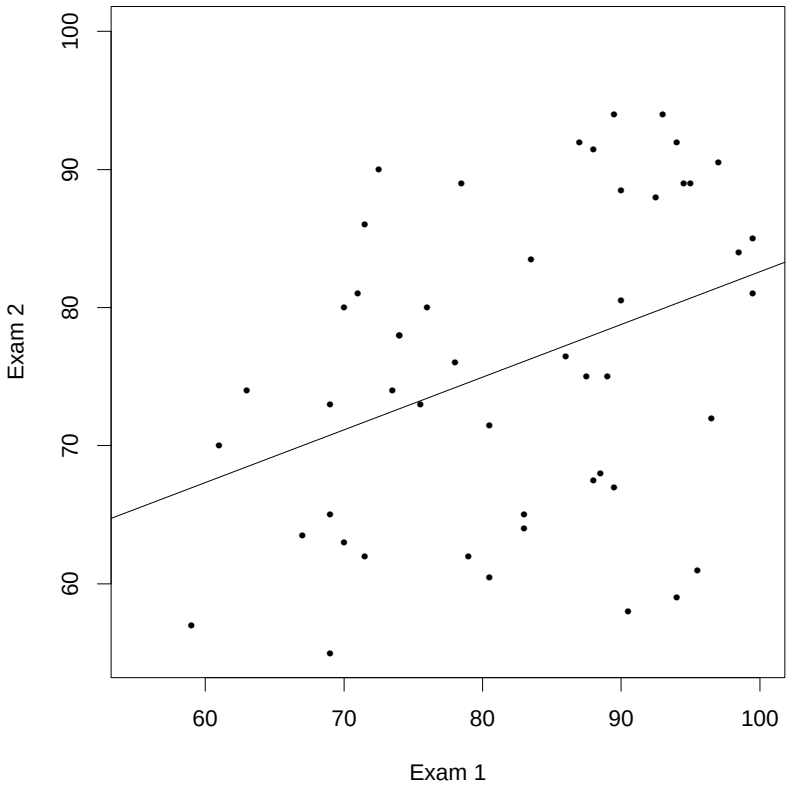

Figure 2: Fitted regression line: $\hat{y} = 44.396 + 0.382x$

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = 2290.73 \quad \sum(x_i - \bar{x})^2 = 5996.445$$

$$\bar{x} = 82.31 \quad \bar{y} = 75.84$$

When Are These Estimates Good?

When certain conditions are met, we can say that our least squares method yields good estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ of β_1 and β_0. These are called the Gauss-Markov conditions.

- $E(\varepsilon_i) = 0$ for all i
- $Var(\varepsilon_i) = \sigma^2$ for all i
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$

These assumptions are also necessary for us to make statements about the mean and variance of the estimates and for further inference about the model parameters.

Measuring the Fit of the Model

Once we've determined our estimated regression line, we'd like to know how well the model fits. How far/close are the observations to the fitted line?

- One way to do this is to see how big the SSE is.
- For simple linear regression $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$ where $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$.
- Note that this quantity depends on the units in which the dependent variable is measured.
- Another way to measure the fit of the model is to look at the proportion of the total variability in the dependent variable that can be explained by the independent variable.

- We can measure the total variability in the dependent variable using the total sum of squares $SST = \sum_{i=1}^{n}(y_i - \bar{y_i})^2$.
- We can measure the variability in the dependent variable that can be explained by the independent variable $SSR = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2$.
- This means that the proportion of total variability in the dependent variable that can be explained by the independent variable is $\frac{SSR}{SST}$.
- This quantity is called the coefficient of determination and denoted by R^2.
- We can prove that $SST = SSR + SSE$. Therefore, R^2 can also be written $\frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST}$.

Sample Correlation

- We'd like to have a way to estimate the true correlation, ρ, using the data.
- This is the sample correlation r, given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- This can be re-expressed in terms that we have used before. Remember, that we can write $\hat{\beta}_1$ as $\frac{S_{xy}}{S_{xx}}$. This yields $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \hat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}}$.
- This relationship also means that we can write the regression equation given r, S_{xx}, S_{yy}, and the sample means of x and y. We know $\hat{\beta}_1 = r\sqrt{\frac{S_{yy}}{S_{xx}}}$.
- In the case of simple linear regression (one independent variable), the coefficient of determination $R^2 = r^2$.

Inferences about the Model Parameters

- The least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained using our sample are only estimates of β_0 and β_1.
- How good are these estimators?
- What are their means, variances, etc.?
- How can we make a confidence interval/hypothesis test for these parameters?

Sampling Distribution for Slope Estimate, $\hat{\beta}_1$ and Intercept Estimate, $\hat{\beta}_0$

- $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_0) = \beta_0$. So $\hat{\beta}_1$ is unbiased for β_1 and $\hat{\beta}_0$ is unbiased for β_0.
- $Var(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nS_{xx}}$ and $Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$ where $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ and $\sigma^2 = Var(Y) = Var(\varepsilon)$.
- $Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}\sigma^2}{S_{xx}}$.
- The distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$ both depend on the distribution of the error term ε. These are normally distributed if ε is normally distributed.
- We will generally be looking at models, in which we assume that ε is normally distributed.

Estimator for σ^2

- We rarely know σ^2, so we will need to estimate it based on the data.
- Since σ^2 represents the variance of the Y_is around the line $\beta_0 + \beta_1 X_i$, it makes sense to estimate it using some function of the distances between the data points and the fitted line.
- This estimator for σ^2 is $S^2 = \frac{SSE}{n-2}$, where $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- S^2 is unbiased for σ^2.

Sampling Distribution of the Parameter Estimators Under Normality

- If the error term ε is normally distributed, then

  1. $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2$ distribution with $n-2$ d.f.

  2. S^2 is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$.

  3. Both $\frac{\hat{\beta}_0-\beta_0}{SE(\hat{\beta}_0)}$ and $\frac{\hat{\beta}_1-\beta_1}{SE(\hat{\beta}_1)}$ have t distribution with $n-2$ d.f. where $SE(\hat{\beta}_0) = S\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}$ and $SE(\hat{\beta}_1) = S\sqrt{\frac{1}{S_{xx}}}$

- Knowledge of the sampling distributions of these statistics enables us to conduct hypothesis tests and form confidence intervals.

Hypothesis Tests/CIs for Coefficients

- After fitting a linear model, we might ask whether there is sufficient evidence to conclude that the x variable is a useful predictor of the y variable.
- This is a hypothesis test with H_0: $\beta_1 = 0$ and H_a: $\beta_1 \neq 0$.
- We can conduct the test as usual, formulating the test statistic as:

$$T = \frac{\hat{\beta}_1 - 0}{S\sqrt{\frac{1}{S_{xx}}}}$$

- Of course, we can also use the same methodology to test hypotheses which involve another value of β_1 (instead of 0) or to test hypotheses involving β_0.

- In general, to test $H_0 : \beta_i = \beta_{i0}$ vs $H_a : \beta_i \overset{>}{\underset{<}{\neq}} \beta_{i0}$ (for $i = 0, 1$) we use the test statistic

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{SE(\hat{\beta}i)}$$

which has a t distribution with $n - 2$ d.f. under H_0.

- Using the information about the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$, we can form confidence intervals for these parameters. To find a $(1 - \alpha)100\%$ confidence interval:

$$\hat{\beta}_i \pm t^{n-2}_{\frac{\alpha}{2}} SE(\hat{\beta}_i)$$

Are the Assumptions of the Model Met?

- Suppose we use least squares to obtain an estimated regression line.
- In order to make inferences concerning the parameters β_0 and β_1, we need to make assumptions about the distribution/correlation of the residuals.
- One way to examine the truthfulness of the assumptions is to look at a scatter plot of the residuals $(e = y - \hat{y})$ vs. the fitted values $(\hat{y})$.
- They should form a cloud (no patterns), symmetric about 0, with fairly even variation of the residuals over the range of fitted values.

Using a log Transform of the Response

- Suppose we suspect an exponential pattern in the data (rather than linear).
- A non-linear model of the form $E(Y) = A\exp(\beta_1 X)$ may be appropriate in this case.
- If we take the log transform of the response/dependent variable, we get a simple linear model: $E[\log(y)] = \log(A) + \beta_1 X = \beta_0 + \beta_1 X$, where the intercept that we're fitting is $\log(A)$.
- We can substitute the estimates that we obtain through the least squares into the original model, yielding $\hat{y} = \exp(\beta_0)\exp(\beta_1 X)$.
- Now, each one-unit change in X means our estimate of Y increases by a factor of e, the base of the natural log.

Confidence Interval for $\mathbf{E(Y)}$

- Remember that our regression line is just an estimate for the expected value of the Y variable.
- This means we're estimating $E(Y) = \beta_0 + \beta_1 x^*$ with $\hat{\beta}_0 + \hat{\beta}_1 x^*$, where x^* is just the value of X for which we want to estimate $E(Y)$.
- We know that since $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators, the quantity $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is an unbiased estimator for $E(Y)$.
- The standard error for our estimate can be shown to be $S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$ where $S^2 = \frac{SSE}{n-2}$.
- This yields a confidence interval for $E(Y) = \beta_0 + \beta_1 x^*$, of the form

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2}}^{n-2} S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Prediction Interval for Y when $\mathbf{X = x^*}$

- Instead of a confidence interval for the mean $E(Y)$, if we want a confidence interval for a prediction of Y when $X = x^*$.

- Before, we were estimating a parameter $E(Y)$. Now we want to estimate the value of a random variable, the Y we observe at some specific time when $X = x^*$.

- Intuitively, we would estimate this value somewhere near the middle of the distribution for Y for $X = x^*$. The center of this distribution is $E(Y) = \beta_0 + \beta_1 x^*$, which is estimated by $\hat{\beta}_0 + \hat{\beta}_1 x^*$.

- So we have the same estimate for $E(Y)$ as we do for a prediction of Y, but intuitively, the variance for a prediction must be larger.

- The SE can be shown to be $S\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}$, yielding a prediction interval for Y (when $X = x^*$ and $S^2 = \frac{SSE}{n-2}$)

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2}}^{n-2} S\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}$$