Stat 111 Spring 2025 Week 11: Simple Linear Regression

1. **Least Squares Fit**
   Suppose we observe pairs of data values x_i and y_i for $i = 1, \ldots, n$ individuals, and that a graph of y vs. x shows a linear association. A linear fit is of the form $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, identifying a 'typical' y value $\hat{y}$ to go with a given x value.

   a) The least squares linear fit chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimizes the sum of squared errors SSE= $\sum(y_i - \hat{y}_i)^2$ for the observed x_i and y_i values. Show that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = r\frac{s_y}{s_x}$. Note the similarity to the conditional mean formula for bivariate Normal variables.

   b) Show the following equality:

   $$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 \;=\; \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \;+\; \sum_{i=1}^{n}((\hat{\beta}_0 + \hat{\beta}_1 x_i - (\beta_0 + \beta_1 x_i))^2$$

   or

   $$\sum(y_i - \mu_i)^2 \;=\; \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \mu_i)^2$$

   Explain how this equality shows, without calculus, that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimates for β_0 and β_1. It will help to show

   $$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i \;=\; \sum_{i=1}^{n} x_i(y_i - \bar{y})$$

   c) Suppose you instead require that $\hat{\beta}_0$ and $\hat{\beta}_1$ be chosen such that $\sum(y_i - \hat{y}_i) = 0$ and $\sum x_i(y_i - \hat{y}_i) = 0$. This forces the fitted mean values ($\hat{y}_i$'s) to have the same average as the y_i's, and for the errors $y_i - \hat{y}_i$ to be uncorrelated with the x_i's. Show that these requirements lead to the same estimates as least squares. Note how neither criterion requires independence or any specific distribution for the errors.

   d) Rearrange the least squares estimates to show the regression towards the mean effect. Explain how this effect applies to both x and y, meaning the least squares fits are not reversible. Consider a hypothetical example to illustrate: the correlation between years of education for husband and wife pairs is $r = 0.5$, with $\bar{x} = \bar{y} = 12$ years and $s_x = s_y = 3$ years.

   e) When setting betting spreads (e.g., for an NFL football game), a casino wants the betting spread x_i to be close to the true spread (home team points minus visiting team points) y_i, with a least squares fit of y_i on x_i that ideally has $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$. Suppose the casino can achieve a correlation of about $r = 0.33$ with the actual spreads, which have a standard deviation of about 15 points. What should be the standard deviation of the betting spreads?

2. **Normal Simple Linear Regression**

   a) The usual simple linear regression model assumes $Y_i = \mu_i + \epsilon_i$, where $\mu_i = \beta_0 + \beta_1 x_i$ and $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, for $i = 1, \ldots, n$. State the four assumptions implied by this equation (a fifth assumption is that the x_i's are fixed and known). Explain how the 1988 OSU blood alcohol experiment data might violate one or more of these assumptions. Subjects choose a number at random between 1 and 9 and drank that many beers. The percentage blood alcohol content (BAC) was measured for each subject by a member of the campus police one hour after drinking.

b) Explain why, for the Normal regression model, the least squares estimates are also maximum likelihood estimates. Find the mle for σ^2.

c) Show that the fitted intercept and slope are unbiased estimates and derive their bivariate Normal joint sampling distribution. Say how the expansion in 1b implies that the mle $\hat{\sigma}^2$ is biased low.

d) Explain how residual plots are useful for checking model assumptions, especially when the model explains much of the variability in the y_i's.

e) Consider modeling the median home selling prices in year 2 as a function of the median price in year 1 for the same municipality. Show that a least squares fit is reasonable both for predicting selling prices and for log-prices, but that log-prices are more appropriate for the Normal regression model. Show the model implied for prices by the Normal linear model on log-prices.

3. **Regression Predictions**

a) Show how to recenter x values to make the intercept into an estimate for the mean Y value for a given x value, and how this provides you with a standard error estimate for this fitted mean value in your regression output. . Demonstrate for data from the 1986 Ohio State blood alcohol experiment. Find a 95% confidence interval for the mean blood alcohol corresponding to $x = 3$ beers. Use $t_{(n-2)}$ percentiles instead of $N(0,1)$ to adjust for having estimating σ^2.

b) Show how to adjust the variance formula for $\hat{\beta}_0$ to account for this shift to give the standard error for a mean response $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Note the quadratic term and explain the implications. Graph blood alcohol against number of beers and draw in the least squares line and the confidence intervals for the mean response.

c) Suppose a new value Y_{n+1} will be observed that is independent of all previous Y_i values. Find the variance of $Y_{n+1} - (\hat{\beta}_0 + \hat{\beta}_1 x_{n+1})$ and explain how this implies a prediction standard deviation formula. Explain the distinction between the prediction interval and the confidence interval for a mean response, and why prediction intervals depend more heavily on the Normal assumption.

d) The James-Stein estimate is useful for estimating multiple means (e.g. for an ANOVA model). Show how the point estimate $\hat{\theta}_i = \hat{B}\bar{y} + (1 - \hat{B})y_i$, with $\hat{B} = \frac{(k-3)V}{\sum(Y_i - \bar{Y})^2}$, approximates the regression of the true group means $\theta_1, \ldots, \theta_k$'s based on $Y_1, \ldots, Y_k$, where $Y_i|\theta_i \overset{\text{indep}}{\sim} N(\theta_i, V)$, with V estimated by MSE$/n$ (assuming equal sample sizes). This is also the posterior mean of $\theta_i|y, V, A$ when. assuming prior distribution $p(\mu, B) \propto 1/B$, for $B = \frac{V}{V+A}$, where A represents the variance of the θ_i's and $Y_i \overset{\text{i.i.d.}}{\sim} N(\mu, V + A)$ is assumed. Show how to compute a posterior standard deviation estimate based on $E(B|y, V)$ and $\text{Var}(B|y, V)$. For example, estimating the $k = 12$ mean scoring rates for WNBA basketball teams. Note how the intervals resemble the regression prediction intervals in that they grow wider as $|x_i - \bar{x}|$ increases.

4. **ANOVA and Lack of Fit**

a) Show that the pooled two-sample t test is a special case of simple regression on an indicator variable. For example, test whether goalkeepers and other players have the same mean height or weight. Show that the regression output for the t-test of $\beta_1 = 0$ exactly matches the pooled 2-sample t test.

b) Show how to represent a 1-way ANOVA model as a Normal linear model with an intercept and $k - 1$ coefficients for indicator variables. Write out the ANOVA table for regression and explain the whole model F test, and how this is the same as the ANOVA F test of equal means in this context. Use the soccer height and weight data as an example.

c) Explain how a simple linear regression model becomes an ANOVA model if you treat distinct x values as categories, rather than as a continuous predictor. Note how you need repeated data values with the same x values in order to fit such a model. Describe the Lack of Fit test for regression and how this *saturated* model is used to find an error estimate that does not depend on the linearity assumption. Demonstrate the lack of fit test for several data sets. Note how failing to reject does not imply linearity, or any other regression assumptions.

d) Consider the simple ANCOVA model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ for $i = 1, \ldots, n$, where x_{1i} is height and x_{2i} is an indicator for GK. Give an expression for the mean weight of goal keepers and for the mean weight of other players, as a function of height. Compare the interpretation of β_1 in this model to β_1 in a simple regression of weight on height.

e) Add $\beta_3 x_{3i}$ to the model in part d, with $x_3 = x_1 x_2$. Explain how this generalizes the model to allow different slopes and intercepts in the two groups.

5. **Fixed Intercept Regression Model**

For estimating the residual variance σ^2, it is easiest to begin with a linear model with a single mean parameter, like our one-sample model. Suppose we know the mean for some value of the single explanatory variable. The x_i's can then be recentered to make $\beta_0 = 0$, so the linear model becomes

$$Y_i \ = \ \beta_1 x_i \ + \ \epsilon_i, \qquad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \qquad i \ = \ 1, \ldots, n$$

For example, with the blood alcohol data, it would not be unreasonable to assume $\beta_0 = 0$. The R command **lm(y$\sim$ 0+x)** fits a no-intercept regression.

a) Find the least squares/maximum likelihood estimate for $\hat{\beta}_1$ and determine its distribution.

b) Show a simpler version of the expansion in presentation 1:

$$\sum((Y_i - x_i\beta_1)^2) \ = \ \sum((Y_i - x_i\hat{\beta}_1)^2) \ + \ (\hat{\beta}_1 - \beta_1)^2 \sum x_i^2$$

c) Show that $\hat{\epsilon}_i \ = \ Y_i - \hat{Y}_i$ and $\hat{Y}_i \ = \ x_i\hat{\beta}_1$ are uncorrelated (so $\hat{\epsilon}_i$ and $\hat{\beta}_1$ are also uncorrelated). These variables have a bivariate Normal joint distribution, so uncorrelated implies independent. Use this fact and the result of part b to show that $\frac{\sum(Y_i - \hat{Y})^2}{\sigma^2} \sim \chi^2_{(n-1)}$, so that $s^2 \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2\sigma^2}\right)$ is an unbiased estimate for σ^2 that is independent of $\hat{\beta}_1$.

d) Use the result of part c to find an expression for a $(1-\alpha)100\%$ t confidence interval for β_1. Give a 95% CI for the increase in mean blood alcohol for each additional beer consumed.

e) If a new subject drinks x_{n+1} beers, find the variance of $Y_{n+1} - x_{n+1}\hat{\beta}$ and use this to find a 95% prediction interval for a person's blood alcohol an hour after drinking 3 beers.