MA477: Data Science
Lesson 7 Outline — 23 January 2026
United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

1 Administrative

- Bonus points
- PSET 1 and 2 Discussion
- Project 1 Discussion
- Student review
- Qualitative Data Exercsise.

2 Qualitative Predictors Lesson Objectives

- Use qualitative predictors in a linear regression model.
- Interpret the coefficient of an interaction term in a linear regression model.
- Employ polynomial regression to accommodate nonlinear relationships

3 Student Review

4 Key Terms

- The **Sum of Square Error** (SSE) is a measure of error for your model to the data $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- The **Sum of Square Total** (SST) is a measure of un-normalized deviation of the data SST $= \sum_{i=1}^{n}(y_i - \bar{y})^2$
- The **Standard Deviation** measures the typical spread of the data around the mean. $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$
- The **The Coefficient of Determination** (R^2) is the percentage of the total observed variation in the response variable that is accounted for by changes in the explanatory variable $R^2 = 1 - \frac{SSE}{SST}$

5 Review: t-Statistic for a Single Regression Coefficient

To test whether a single predictor has a linear association with the response, we test the null hypothesis

$$H_0 : \beta_j = 0$$

against the alternative

$$H_a : \beta_j \neq 0.$$

After fitting the linear regression model using least squares, we obtain:

- $\hat{\beta}_j$, the estimated coefficient for predictor j,
- SE($\hat{\beta}_j$), the standard error of $\hat{\beta}_j$, which estimates the standard deviation of its sampling distribution.

The **t-statistic** is defined as

$$t_j = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}.$$

Under the null hypothesis and the linear model assumptions, t_j follows a t-distribution with $n - p - 1$ degrees of freedom, where n is the sample size and p is the number of predictors. The corresponding p-value is computed from this distribution and used to determine whether to reject H_0.

6 Review: F-Statistic for Multiple Regression Coefficients

To test whether the predictors are jointly associated with the response, we test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

against the alternative that at least one coefficient is nonzero.

Define:

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad \text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where TSS is the total sum of squares and RSS is the residual sum of squares from the fitted model.

The **F-statistic** is

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}.$$

Under the null hypothesis and the linear model assumptions, F follows an F-distribution with p and $n-p-1$ degrees of freedom. Large values of F provide evidence against H_0.

Partial F-Test for a Subset of Coefficients

Sometimes we wish to test whether a subset of q coefficients is equal to zero:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0.$$

We fit:

- a **restricted model**, excluding these q predictors, with residual sum of squares RSS_0,
- the **full model**, including all predictors, with residual sum of squares RSS.

The corresponding **partial F-statistic** is

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)}.$$

This statistic tests whether the excluded predictors jointly improve model fit beyond what would be expected due to random variation alone.

7 Three Perspectives on Linear Regression

Three Perspectives on Linear Regression (Statistical Error focus)

Let $X \in \mathbb{R}^{n\times p}$ be the design matrix, $\mathbf{y} \in \mathbb{R}^n$ the response vector, and $\boldsymbol{\beta} \in \mathbb{R}^p$ the coefficient vector.

—

Statistical (Probabilistic) Perspective and Maximum Likelihood

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

The observed response is modeled as a deterministic linear component plus random noise.

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

The noise is assumed independent, mean zero, and normally distributed with constant variance.

$$p(\mathbf{y} \mid X, \boldsymbol{\beta}) \;\propto\; \exp\!\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2\right)$$

Under the Gaussian assumption, the probability of the data decreases as squared error increases.

$$\ell(\boldsymbol{\beta}) = \log p(\mathbf{y} \mid X, \boldsymbol{\beta}) \ \propto \ -\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$$

Maximizing the log-likelihood is equivalent to minimizing squared prediction error.

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$$

The maximum likelihood estimator selects coefficients that make the observed data most probable.

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

Solving the likelihood maximization yields the familiar least-squares estimator.

Interpretation: Linear regression estimates the conditional mean

$$\mathbb{E}[\mathbf{y} \mid X] = X\boldsymbol{\beta},$$

and least squares arises naturally as the maximum likelihood estimator when errors are Gaussian.

Summary

- **Geometric:** projection of $\mathbf{y}$ onto $\mathrm{Col}(X)$
- **Statistical:** maximum likelihood under a Gaussian noise model
- **Optimization:** minimization of squared prediction error

All three perspectives describe the same estimator $\hat{\boldsymbol{\beta}}$ from different viewpoints.

Discussion Questions

1. How do we integrate categorical variables into linear models?
2. Compare and contrast the left model to the right model - what does the interaction term do? The left model includes income and a student indicator but no interaction:

   $$\text{balance}_i = \beta_0 + \beta_1 \,\text{income}_i + \beta_2 \,\text{student}_i + \varepsilon_i.$$

   This specification produces two parallel regression lines. Non-students have slope β_1 and intercept β_0, while students have the same slope β_1 and a shifted intercept $\beta_0 + \beta_2$. Thus, the marginal effect of income on balance is assumed to be the same for both groups.

   The right model includes an interaction between income and student status:

   $$\text{balance}_i = \beta_0 + \beta_1 \,\text{income}_i + \beta_2 \,\text{student}_i + \beta_3 \,(\text{income}_i \times \text{student}_i) + \varepsilon_i.$$

   In this model, non-students have slope β_1, while students have slope $\beta_1 + \beta_3$. The interaction coefficient β_3 is the difference in slopes between students and non-students, allowing the effect of income on balance to vary by student status.
3. Interpret the regression coefficient for income vs. debt considering the interaction term of being a student. - note that the slope for students is lower than the slope for non-students. This suggests that increases in income are associated with smaller increases in credit card balance among students as compared to non-students
4. Why is a polynomial regression model still considered a linear model? - it is linear in its parameters, even though it is nonlinear in the predictor. The response is expressed as a linear combination of unknown coefficients multiplying known transformations of the predictor, such as powers of the predictor. As a result, the model can be estimated using ordinary least squares and falls within the class of linear regression models.
5. What are the risks of using a very high-degree polynomial in regression?

98 3. Linear Regression

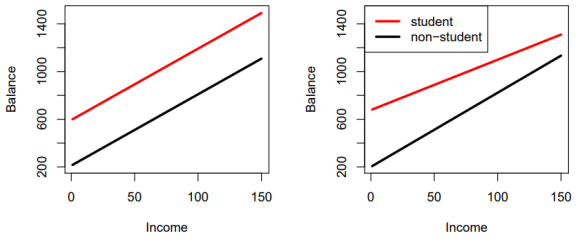

FIGURE 3.7. *For the* `Credit` *data, the least squares lines are shown for prediction of* `balance` *from* `income` *for students and non-students.* Left: *The model (3.34) was fit. There is no interaction between* `income` *and* `student`. Right: *The model (3.35) was fit. There is an interaction term between* `income` *and* `student`.

Figure 1: Interaction Term

1. Qualitative Predictors in Linear Regression

Linear regression can incorporate qualitative (categorical) predictors by encoding them using **dummy variables**. Each dummy variable takes values 0 or 1 to indicate group membership.

Binary Qualitative Predictor

Let X be a binary variable:

$$X = \begin{cases} 1 & \text{if Female} \\ 0 & \text{if Male} \end{cases}$$

The regression model is:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Interpretation:

- β_0: mean response for the baseline group (Male)
- β_1: difference in mean response between Female and Male

The coefficient β_1 measures how the expected response changes when moving from the reference category to the indicated category.

Qualitative Predictors with Multiple Levels

A qualitative predictor with K categories is represented using $K - 1$ dummy variables. One category is omitted and serves as the **reference group**.

Example with three regions:

$$Y = \beta_0 + \beta_1 X_{\text{South}} + \beta_2 X_{\text{West}} + \varepsilon$$

Each coefficient compares its category to the reference category.

Key rule: Including all K dummy variables causes perfect multicollinearity and must be avoided.

—

2. Interaction Terms in Linear Regression

Interaction terms allow the effect of one predictor to depend on the value of another predictor.

Continuous × Binary Interaction

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \cdot D) + \varepsilon$$

where:

- X is a continuous predictor
- D is a binary indicator variable

Interpretation:

- β_1: slope of X for the baseline group ($D = 0$)
- β_3: change in the slope of X when $D = 1$
- Slope for $D = 1$: $\beta_1 + \beta_3$

Without an interaction term, the model assumes the relationship between X and Y is the same across groups.

—

3. Polynomial Regression for Nonlinear Relationships

Polynomial regression extends linear regression by including powers of a predictor.

Quadratic Model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Although the relationship between X and Y is nonlinear, the model remains linear in the coefficients.

Higher-Order Polynomials

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + \varepsilon$$

Trade-offs:

- Higher degree increases flexibility
- Excessive degree can lead to overfitting and unstable predictions

Model complexity should be guided by diagnostics or validation rather than fit alone.

—

Conceptual Summary

- Qualitative predictors model group-level differences
- Interaction terms allow relationships to vary across groups
- Polynomial regression captures smooth nonlinear patterns

Central idea: Linear regression is flexible not because predictors must be linear, but because the model is linear in its parameters.